ESD-TR-80-90

(12)

Technical Report 541

See 1473.

Speech Transformation System (Spectrum and/or Excitation) Without Pitch Extraction

S. Seneff

31 July 1980

Prepared for the Defense Advanced Research Projects Agency
under Electronic Systems Division Contract F19628-80-C-0002 by

Lincoln Laboratory

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

LEXINGTON, MASSACHUSETTS

The work reported in this document was performed at Lincoln Laboratory, a center for research operated by Massachusetts Institute of Technology. This work was sponsored by the Defense Advanced Research Projects Agency under Air Force Contract F19628-80-C-0002 (ARPA Order 3673).

The views and conclusions contained in this document are those of the contractor and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the United States Government.

This technical report has been reviewed and is approved for publication.

FOR THE COMMANDER

Raymond L. Loiselle, Lt. Col., USAF
Chief, ESD Lincoln Laboratory Project Office

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
LINCOLN LABORATORY

SPEECH TRANSFORMATION SYSTEM (SPECTRUM AND/OR EXCITATION) WITHOUT PITCH EXTRACTION

S. SENEFF
Group 24

TECHNICAL REPORT 541

31 JULY 1980

LEXINGTON MASSACHUSETTS

ABSTRACT*

A new speech analysis-synthesis system has been developed which is capable of independent manipulation of the fundamental frequency and spectral envelope of a speech waveform. The system deconvolves the original speech with the spectral-envelope estimate to obtain a model for the excitation. Hence, explicit pitch extraction is not required. As a consequence, the transformed speech is more natural sounding than would be the case if the excitation were modeled as a sequence of pulses. The system has applications in the areas of voice modification, baseband-excited vocoders, time-scale modification, and frequency compression as an aid to the partially deaf.

* This report is based on a thesis of the same title submitted to the Department of Electrical Engineering and Computer Science at the Massachusetts Institute of Technology, on 10 May 1980, in partial fulfillment of the requirements for the degrees of Master of Science and Electrical Engineer.

CONTENTS

ACKNOWLEDGMENTS

I wish first to thank Dr. Victor Zue for an inspiring lecture in a course on speech processing, during which the idea for this report was born. Next, I thank Dr. Ben Gold for many contemplative blackboard discussions, particularly on the digital-signal-processing aspects of the problem. Finally, I am grateful to Professor Ken Stevens for insightful remarks on issues related to production and perception of speech, and for time spent poring over early drafts of my thesis.

SPEECH TRANSFORMATION SYSTEM (SPECTRUM AND/OR EXCITATION) WITHOUT PITCH EXTRACTION

I. INTRODUCTION

A steady-state speech waveform can generally be modeled as a convolution of two distinct components, one of which is loosely referred to as the excitation and the other as the vocal-tract filter. During voiced utterances, the excitation consists of a quasi-periodic series of pulses produced by the vibrations of the vocal folds. The shape of the vocal tract, controlled by the positioning of the articulators (tongue, jaws, etc.) determines the frequency response of the vocal-tract filter. Unvoiced sounds are usually produced by a tight constriction at some point along the vocal tract. Again the vocal-tract shape, especially to the front of the constriction, determines the frequency shaping, but the excitation is a sound pressure noise source with no periodicities evident.

To construct a simplified digital model of the speech waveform, a common approach is to model the excitation as a sequence of unit samples spaced by the fundamental period during voiced utterances, or as a pseudorandom noise function during unvoiced utterances. The spectral envelope is then modeled as a frequency-shaping filter, which incorporates the formant resonances, the high-frequency falloff due to the fact that glottal pulses are not really impulses, as well as the radiation effects. When the simplified excitation is processed through a carefully selected frequency-shaping filter, a sound is produced which closely resembles true speech (see Ref. 1, pp. 379-385).

A vocoder, or voice encoder, is an analysis-synthesis system which characterizes a true speech utterance by a set of parameters describing the excitation and the frequency-shaping filter as a function of time (see Ref. 1, pp. 324-334). Typically, a vocoder models the excitation as either voiced or unvoiced, and during voiced segments of speech extracts an estimate of the fundamental frequency. The frequency-shaping filter is also estimated by some form of spectral analysis of a small segment of the waveform. Since the short-time spectrum of the speech is not in fact stationary, but rather varies slowly with time, the parameters must be updated frequently, usually on the order of every 10 to 20 ms.

The synthesis is achieved by convolving the modeled excitation with the filter specified by the spectral parameters. Generally, some sort of parameter interpolation is done to assure a smooth transition whenever a new set of parameters is provided.

Once the parameters describing the simplified excitation and spectral envelope have been derived, it is possible to perform certain transformations on these parameters to produce a modified speech waveform. For example, the apparent pitch of the speaker can be altered by multiplying the derived estimate for the fundamental frequency by a fixed factor prior to the reconstruction of the synthetic speech.

The temporal characteristics of the speech can also be altered, by updating the parameters at the synthesizer at a different rate from the rate of extraction of the parameters at the analyzer. For example, if the speech were sampled at a 10-kHz sampling rate, then 100 samples of input speech would correspond to 10 ms. If it is assumed that the analyzer updates the parameters every 10 ms (i.e., has a 10-ms <u>frame</u> rate), then the synthesizer could simply generate 200 rather than 100 samples prior to the introduction of each new set of parameters. After the entire utterance is processed, the synthesized speech contains twice as many samples as the analyzed speech. Played out at the same clock rate, i.e., 10-kHz sampling rate, the synthesized speech would be twice as slow as the original.

The above alteration of temporal characteristics can be transformed to an alteration of spectral characteristics by playing out the synthesized speech at a different sampling rate. In the above example, if the D/A clock rate is set to twice the A/D clock rate, then the synthesized speech will be frequency-expanded, rather than time-expanded, by a factor of 2.

It is possible to preserve the fundamental frequency in conjunction with an alteration in the spectral envelope, by premultiplying the extracted fundamental period by the correct factor prior to reconstruction of the synthetic speech. In the above system, a doubling of the fundamental period will cause twice as many samples to be generated per pitch period. Thus, when the D/A clock rate is also doubled relative to the A/D clock rate, each pitch period will be restored to its original time interval. Hence, there is independent control over the fundamental-frequency manipulations and the spectral/time manipulations.

A vocoder can thus be used to generate a variety of interesting transformations on the speech waveform. Two possible applications, in addition to time modification, are voice modification and frequency compression. To convert a male voice into a female-like voice, an appropriate first-order approximation is to expand the vocal-tract spectrum by 20 percent and double the fundamental frequency. To realize these specifications the extracted fundamental frequency is multiplied by 2/1.2, 120 output samples are generated for each 100 input samples, and the synthetic speech is played out at a 20-percent increased clock rate.

Another possibility is to use the vocoder to compress the spectrum into low frequencies to bring the information down into a range that is potentially usable by people with high-frequency hearing loss.[2] It is likely that, at least for male voices, a more speech-like result would be realized if the fundamental frequency were left unchanged, while the spectral envelope was compressed. An analogous transformation is needed to restore helium speech spoken by divers.[3] With a vocoder, a spectral compression by a factor c, with no alteration of the fundamental frequency, could be achieved by multiplying the extracted fundamental frequency by c, generating 1/c of the original number of samples per frame, and playing out the synthetic speech at 1/c of the original sampling rate.

Such transformations are relatively straightforward to realize, given a vocoder which represents the voiced excitation as a sequence of unit samples spaced by the predetermined fundamental period. Unfortunately, the quality of the vocoded speech is critically dependent upon the pitch-extraction algorithm, and pitch errors (which are nearly inevitable) produce intrusive glitches in the synthetic speech. Furthermore, the vocoded speech has a distinct synthetic quality, due to the simplified excitation model.

Another approach to obtaining a model for the excitation is to process the waveform with a filter whose response is the inverse of the spectral envelope. The resulting signal e(n), as shown in Fig. 1, has a flat spectrum envelope and is periodic with the frequency of the glottal excitation. If it could be modified in some way to effect a change in the fundamental frequency, then the modified excitation function could be used to excite the vocal-tract filter, without ever resorting to an explicit extraction of the fundamental frequency. The expectation is that synthetic speech using the spectrally flattened waveform as excitation would sound smoother and more natural than vocoded speech.

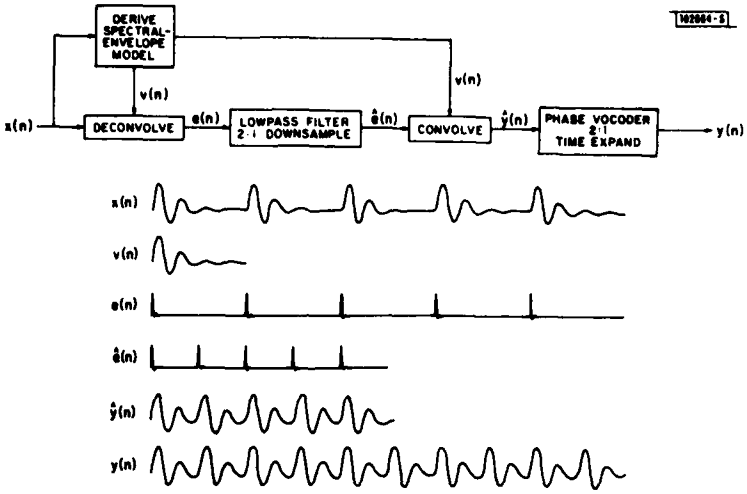
102884-S
DERIVE SPECTRAL-ENVELOPE MODEL
v(n)
v(n)
x(n)
DECONVOLVE
e(n)
LOWPASS FILTER 2:1 DOWNSAMPLE
ê(n)
CONVOLVE
ŷ(n)
PHASE VOCODER 2:1 TIME EXPAND
y(n)
x(n)
v(n)
e(n)
ê(n)
ŷ(n)
y(n)

Fig. 1. A method for doubling fundamental frequency without altering spectral envelope.

However, when the excitation is obtained by deconvolution, there is no longer the capability of controlling the number of samples generated per frame. In order to shorten the fundamental period by a factor of 2, it is necessary to lowpass-filter and 2:1-downsample the derived excitation function prior to the introduction of the frequency-shaping filter. This step is illustrated in Fig. 1, where the resulting downsampled error signal is labeled $\hat{e}(n)$. The number of synthetic speech samples derived from $\hat{e}(n)$ is automatically also reduced by the factor of 2, relative to the original x(n). Hence, speech is generated [$\hat{y}(n)$ in the figure] which, if played out at the original clock rate, will contain the correct frequency-shaping characteristics, and a fundamental frequency which is twice the original value. Unfortunately, the speaking rate will also, necessarily, be twice as fast.

The correct speaking rate can be restored by processing the synthetic speech through a phase vocoder.[4,5] A phase vocoder obtains a high-resolution spectrum every few milliseconds, and transforms the spectral information to polar coordinates, to obtain a magnitude and phase spectrum. Time/frequency characteristics can be modified by multiplying the (unwrapped) phase spectrum by the appropriate constant. (Phase must be unwrapped to remove 2π discontinuities, as will be discussed later.)

The unwrapped phase measured from a single narrow-bandpass filter corresponding to one spectral coefficient is approximately the integral in time of the frequency of the harmonic passed by that filter (see Sec. II). Thus, when the phase is multiplied by a factor c, the result is equivalent to multiplying the frequency passed by the filter by the same factor. The resulting frequency-scale modification can be converted to time-scale modification by altering the D/A clock rate. With the appropriate choice of the factor c and the D/A clock rate, a specified adjustment of both frequency and temporal characteristics can be achieved.

To recapitulate, a possible system for altering the fundamental frequency without changing either the resonance or temporal characteristics is as illustrated in Fig. 1. A model for the spectral-envelope filter v(n) is derived from a small portion of the input waveform x(n). Deconvolution of x(n) with v(n) yields the error signal e(n) in the figure. This error signal can then be lowpass-filtered and 2:1-downsampled to yield a new error signal $\hat{e}(n)$ which, if played out at the original clock rate, would generate a fundamental frequency that is twice as high as the original fundamental frequency. The modified error signal is then processed back through the spectral-envelope filter, and the output $\hat{y}(n)$ is processed through a phase vocoder to 2:1 time-expand the signal. Finally, y(n) is played out with a D/A clock rate equal to the original A/D clock rate. As shown in the waveform diagrams below the block diagram in Fig. 1, y(n) has the same number of samples as x(n) but glottal pulses are occurring twice as often, as desired.

If it is desired to compress the spectrum by some amount while preserving the fundamental frequency intact, it is not necessary to process the synthesized speech through a phase vocoder. In this case, the downsampled waveform is played back at a lower sampling rate and, therefore, temporal characteristics are preserved. In the example of Fig. 1, if $\hat{y}(n)$ were played out at half the original sampling rate, time would be preserved, but frequencies would be compressed by a factor of 2. Thus, applications such as frequency compression as an aid for persons suffering from high-frequency hearing loss, or the restoration of divers' speech, do not require the additional processing through the phase vocoder, and hence are more straightforward than applications requiring additional time/frequency-scale modification.

Lowering the fundamental frequency is an intrinsically more difficult problem than raising it, because it is necessary to generate additional harmonics rather than to remove higher harmonics. The problem is similar to that of high-frequency regeneration from baseband excitation – a problem which has been addressed by several researchers.[6-8]

The goal of these researchers was to reduce the required bit rate for speech transmission by deriving a major portion of the excitation spectrum from the low-frequency information. The two basic approaches to generation of higher harmonics from the lower ones are either to apply nonlinearities to the signal[9] or to copy-up the existing harmonics into the higher frequencies simply by duplicating the spectrum.[7] The major disadvantage of the nonlinearity method is that interharmonic noise tends to be enhanced relative to the signal by the nonlinear process. With the duplication technique, the higher harmonics are not necessarily located at integer multiples of the fundamental. Particularly for high-pitched voices, annoying background pitches are at times perceived superimposed on the voice pitch.

In lowering the fundamental frequency, by whatever technique, temporal characteristics again are destroyed. For example, a simple mechanism is to 2:1-upsample the error signal (derived by convolving the waveform with the inverse filter of the spectral envelope) alternating zero samples with data samples. Since the upsample step is not followed by a lowpass filter, the upper half of the spectrum is a folded replica of the first half. If the upsampled data set is reshaped with the spectral-envelope filter and then played out at the original clock rate, the effect will be to preserve the spectrum, lower the fundamental frequency by a factor of 2, and time-expand the utterance by a factor of 2. Additional processing through a phase vocoder is then necessary to restore the correct temporal characteristics.

The goal of this report is to develop and implement a system which is capable of independent manipulation of the excitation and spectral-envelope characteristics without resorting to an explicit parametric representation of the fundamental-frequency contour. The major issues addressed are:

(a) Selection of an appropriate model for the spectral-envelope filter.

(b) Development of a method for generating higher harmonics from those available in order to minimize the effects of frequency-offset problems and interharmonic noise enhancement, and

(c) Integration of the system with a phase vocoder to restore desired temporal characteristics.

Derivation of the spectral envelope, harmonic duplication, and time/frequency-scale modification are the basic tools needed to realize a large battery of potentially useful transformations of the speech waveform. Among the applications discussed here are frequency compression, baseband-excited vocoding, voice transformation, and time-scale modification.

Section II is an overview of the structure of a system which was developed to meet the above design goals. Following the overview, various aspects of the system are dealt with in detail, in conjunction with the description of specific transformations that were implemented. Next, some results are presented, in the form of waveforms, spectra, and spectrograms, illustrating the various transformations. Finally, suggestions are made for possible further research both in improving the efficiency of computation or quality of the transformed speech, and in further extensions of the applications.

II. GENERALIZED SYSTEM STRUCTURE

In this section we describe the basic generalized system used to perform various transformations on the speech waveform. A detailed description of the individual components, as well as the specific system design related to distinct transformations, will be presented in subsequent sections.

As described in Sec. I, the first step in Fig. 1 is the derivation of the spectral envelope. Many of the techniques available for extraction of the spectral envelope begin with some form of spectral analysis. Since the phase vocoder, needed in the final step to restore temporal characteristics, also requires spectral analysis, it seems logical to attempt a restructuring of the system to combine the two spectral analysis steps into one.

Figure 2(a) is a block diagram of the system which was implemented, motivated by the above arguments. For clarity, the example is the specific application already discussed in Fig. 1, namely a 2:1 increase in the fundamental frequency without alteration of the spectral envelope. Later, a more generalized version will be described.

For this system structure, all the deconvolution, convolution, and filtering steps are done in the frequency domain. The return to the time domain is accomplished in the final step, by summing the modified spectral coefficients. The system operates at the sampling rate, hence each box is updated with each new input speech sample. The system is therefore computationally expensive. It is possible to reconfigure the system to take advantage of the FFT-type phase-vocoder algorithm as implemented by Portnoff,[4] at considerable savings in computational time, as will be described later.

Returning to Fig. 2(a), we observe that the first step is to compute a high-resolution spectrum of a windowed portion of the input waveform. The window should be long enough to include a few pitch periods, but short enough such that the spectral envelope can be approximated as stationary. A good choice is thus about 30 ms.

Once the spectrum has been computed, it is converted from rectangular to polar coordinates to yield a magnitude spectrum and a phase spectrum. The spectral envelope is derived from the magnitude spectrum by smoothing in frequency, to remove the pitch ripple, as shown in Fig. 2(b), (I) and (II).

The time-domain deconvolution to obtain the excitation magnitude spectrum is achieved by dividing the magnitude spectrum by the envelope, yielding the spectrum shown schematically in Fig. 2(b), (III). The lowpass filter [part (IV) of the figure] with cutoff frequency $\pi/2$ is realized by simply discarding the upper half of the flattened magnitude and phase spectra.

The lower half of the magnitude spectrum must now be reshaped with the spectral-envelope data [part (II) in Fig. 2(b)]. Since the upper half of the magnitude spectrum has been discarded, it contains only half as many samples as does the spectral-envelope spectrum. Hence, the envelope spectrum must be 2:1-downsampled in frequency, as shown in part (V) of Fig. 2(b), prior to the sample-by-sample multiplication that is needed to reshape the excitation magnitude spectrum with the spectral envelope [part (VI)].

The final frequency-domain transformation step is to remap the reshaped magnitude spectrum from 0 to $\pi/2$ into the entire frequency band. This step is achieved by modifying the phase

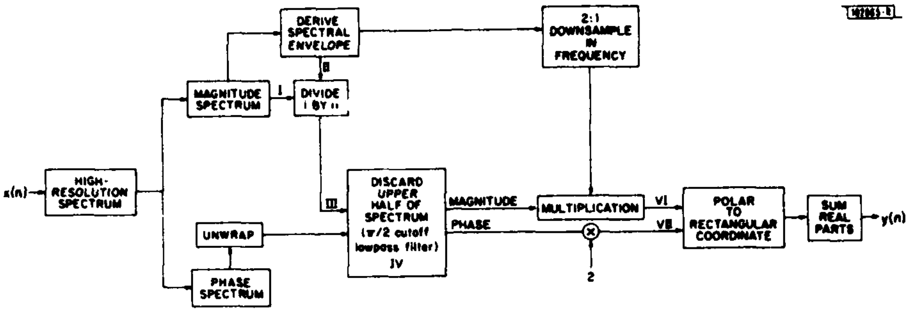

Fig. 2(a). Rearrangement of Fig. 1 such that phase vocoder is integrated into rest of system. This system operates entirely in frequency domain. Time expansion is achieved by doubling unwrapped phase component of spectrum.

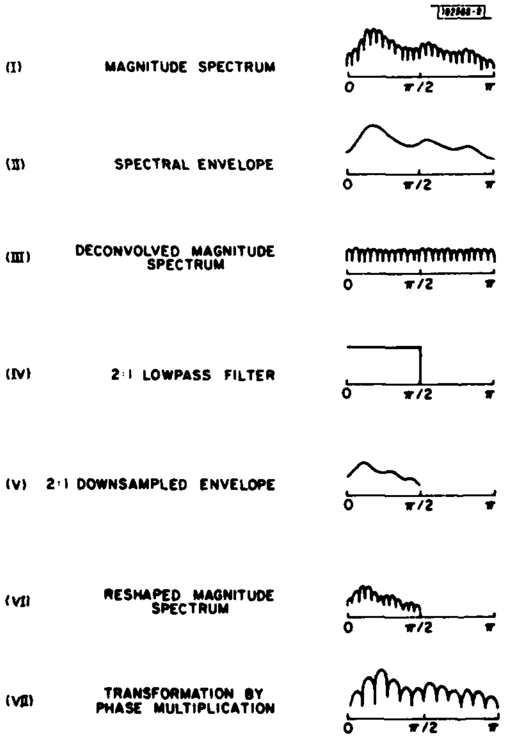

Fig. 2(b). Schematic spectra corresponding to (I) through (VII) in Fig. 2(a).

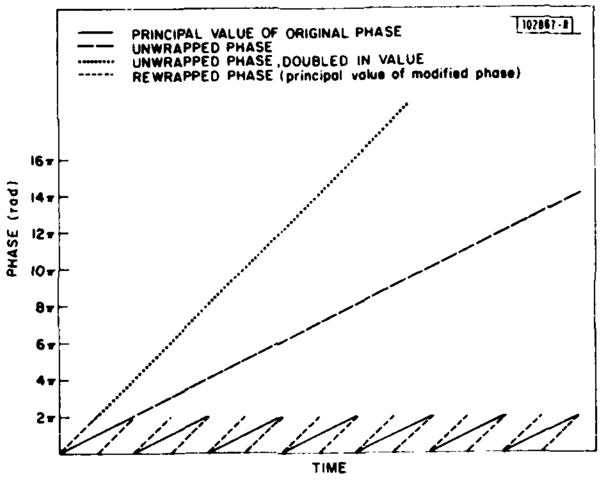

Fig. 3. Illustration of how doubling unwrapped phase of time waveform passed by a single narrowband filter (one DFT coefficient) is equivalent to doubling frequency of sine wave passed by that filter.

spectrum, as outlined in Fig. 3. The unwrapped phase component, measured as a function of time, from a single narrowband filter (corresponding to one DFT coefficient) is approximately a straight line with slope equal to the frequency of the harmonic passed by that filter, as shown in the figure. If this component is then doubled, the slope of the line (and thus also the frequency of the harmonic present in that filter) is also doubled. When the phase component of each DFT coefficient is doubled, the total effect is to double the frequencies of all the harmonics present in the original waveform, thus, effectively, expanding the spectral scale by a factor of 2. Hence, the phase modification achieves the transformation of the magnitude spectrum from item (VI) in Fig. 2(b) to item (VII), a spectrum which spans the entire available band. The new magnitude spectrum has the correct spectral envelope, but only contains half as many harmonics of the fundamental frequency. Hence, the fundamental frequency has been doubled, as desired [Fig. 2(b), (VII)].

The final step is to return to the time domain. The modified magnitude and phase spectra must first be converted back to rectangular coordinates. The real parts are then summed, as will be described in Sec. III, to yield the output sample y(n).

A more general version of the above system is given in Fig. 4. The high-resolution spectrum is actually realized by an iterative technique using a bank of frequency-sampling filters, as will be described in Sec. III. The hanning window is achieved by convolution in frequency rather than the usual multiplication in time, because the iteration procedure precludes the latter possibility.

The spectral envelope is derived by linear filtering of the magnitude spectrum, as will be described further in Sec. IV. A divide, sample-by-sample in frequency, achieves the deconvolution in time. For voiced sounds, the resulting flattened spectrum will have a harmonic structure due to the periodic source. Unvoiced sounds will yield a flattened noise source spectrum. The phase spectrum scale factor c is not restricted to the integer 2, but can be any integer or fraction. Its actual value depends upon the specific application. The box labeled "foreshorten or duplicate" corresponds to the lowpass-filter step of Fig. 2(a). For some applications, additional

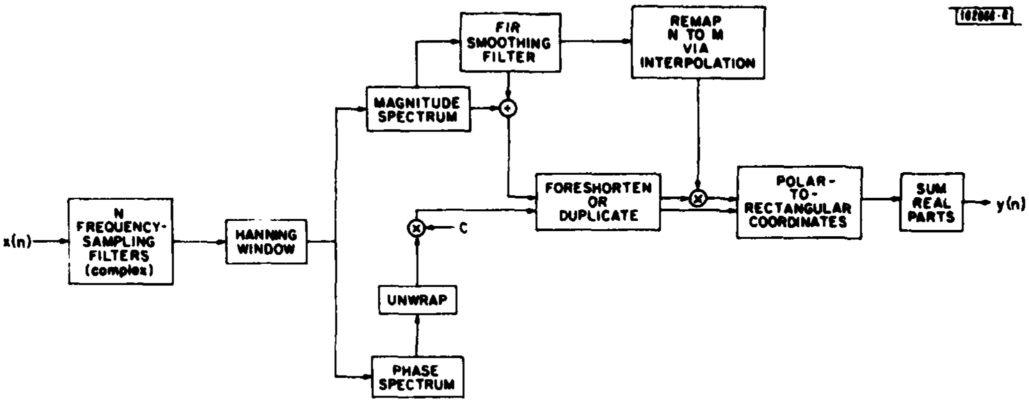

Fig. 4. Extension of Fig. 2(a) to general case where both spectral envelope and fundamental frequency may be scaled by noninteger amounts. Step "foreshorten or duplicate" will become clearer later in text.

harmonics must be generated from the available ones. A discussion of the technique used will be deferred until Sec. V.

The number of samples characterizing the spectral envelope must be changed from the original N to a new number M equal to the number of samples available in the modified excitation magnitude spectrum. In general, this remapping will necessitate some interpolation between the available samples, although, in the specific example of Fig. 2(a), a simple 2:1-downsample was adequate. The final step to recover y(n) by adding real parts is unchanged from the previous example.

Notice that the phase-multiplication step precedes the "foreshorten or duplicate" step in Fig. 4. For the example of Fig. 2(a), the order in which these two steps are done makes no difference, but for applications involving duplication of the phase spectrum, the multiplication step should be done first.

The system was implemented on the Lincoln Digital Speech Processor (LDSP), a high-speed microprocessor with a 50-ns cycle time.[10] All computations were done using fixed-point arithmetic, double precision being used only for the DFT calculation. Analog speech was pre-emphasized, lowpass-filtered with a cutoff frequency of 4kHz, and sampled at an 8-kHz A/D clock rate. A 256-point DFT was computed at the sampling rate, corresponding to a time window of 32 ms, using an iterative method akin to frequency-sampling filters.

The system was developed without regard to computational efficiency, because we felt that implementation of structures such as the Portnoff FFT phase vocoder[4] would make an already complex structure conceptually intractable. However, once it has been demonstrated that the system is capable of generating superior-quality speech, it should be possible to reconfigure it using an FFT-type phase vocoder, at considerable savings in computational requirements, as will be discussed in Sec. VIII.

III. ITERATIVE DISCRETE FOURIER TRANSFORM

An N-point DFT can be viewed as a filter bank containing N/2 equally spaced filters spanning the frequency band from 0 to π, as well as a symmetric set covering the corresponding negative frequencies. The expression for the DFT coefficient $X_k(m)$ of a segment of speech including the time samples from $(m - N + 1)$ to m is

$$X_k(m) = \sum_{n=m-N+1}^{m} x(n)\, W^{k(n-m+N-1)} \quad ; \quad W = e^{j2\pi/N} \tag{1}$$

where k is a frequency index, and m is a time index. This equation can be interpreted as an FIR filter, where the $X_k(m)$ are the output waveform, and the W^i are the filter coefficients.

The corresponding equation for the DFT coefficient $X_k(m - 1)$, evaluated at time $m - 1$ is

$$X_k(m - 1) = \sum_{n=m-N}^{m-1} x(n)\, W^{k(n-m+N)} \quad . \tag{2}$$

In comparing Eq. (1) with Eq. (2), it becomes evident that the summations are over all of the same samples except that the first sample $x(m - N)$ has been dropped off the end, and the new sample $x(m)$ has been added. Furthermore, the coefficients differ by a constant factor W^{-k}. This result is expected, since the input waveform at time m is a time shift by one sample of the input waveform at time $m - 1$, with the insertion of the new sample $x(m)$ in place of the circularly shifted oldest sample $x(m - N)$. The circular rotation in time is equivalent to the multiplication of each frequency coefficient by W^{-k}. Since the time shift can be expressed as a multiplication factor in frequency, it should be possible to derive each DFT coefficient at time m from its value at time $m - 1$, to produce an iterative process.

Mathematically,

$$X_k(m) = \sum_{n=m-N}^{m-1} x(n)\, W^{k(n-m+N)}\, W^{-k} + x(n)\, W^{k(N-1)} - x(n)\, W^{-k} \quad .$$

Since $W^{Nk} = 1$, this reduces to

$$X_k(m) = W^{-k} X_k(m - 1) + W^{-k}[x(m) - x(m - N)] \quad .$$

Thus, each DFT coefficient $X_k(m)$ is the output of a recursive filter with a frequency response as follows:

$$H_k(z) = \frac{W^{-k}(1 - z^{-N})}{1 - W^{-k} z^{-1}} \quad .$$

This filter is a specific application of the generalized formula described by Oppenheim and Schafer[11] for frequency-sampling filters. It contains N equally spaced zeros around the unit circle and a single pole at $z = W^k$, which cancels the zero at frequency $2\pi k/N$. Hence, it is a complex bandpass filter centered at the pole frequency.

Two issues in the implementation of the DFT iteratively are the question of stability and the introduction of a non-square window. Since the pole lies on the unit circle, the system may well become unstable due to computational errors. The stability problem can be solved by evaluating the z-transform on a circle of radius slightly greater than one, as follows:

$$z' = z/r \quad ; \quad r \lesssim 1$$

$$H'_k(z) = H_k(z')\Big|_{z'=z/r} = \frac{W^{-k}(1 - r^N z^{-N})}{1 - W^{-k} r z^{-1}} \quad .$$

This procedure effectively brings both the zeros and the pole slightly inside the unit circle.

Clearly, it is preferable to taper the waveform by a window such as the hanning window (raised cosine) in order to improve the frequency characteristics of each filter output. This windowing cannot be done in time in conjunction with the iteration, since all the window coefficients would change at each increment. The solution is to implement the window by convolution in frequency of the filter outputs with the transform of the window. Fortunately, the transform of a raised cosine is a very simple function, with only three nonzero values:

$$hw(n) = \frac{1}{2}\,(1 - \cos 2\pi n/N)$$

$$= \frac{1}{2}\,[1 - \frac{1}{2}(e^{j2\pi n/N} + e^{-j2\pi n/N})]$$

$$HW(k) = \frac{1}{2}\,u_o(k) - \frac{1}{4}[u_o(k+1) + u_o(k-1)] \quad .$$

If the DFT is computed at the sampling rate, only one sample of the inverse DFT needs to be computed in order to recover the speech waveform. This sample should be the one in the middle of the cosine window, i.e., the sample at time $[m - (N/2)]$:

$$x(m - \frac{N}{2}) = \frac{1}{N}\sum_{k=0}^{N-1} X_k(m)\, W^{[m-(N/2)]\,k} = \frac{1}{N}\sum_{k=0}^{N-1} X_k(m)\, e^{j\pi k} \quad .$$

Since $e^{j\pi k}$ is simply ±1, this calculation reduces to adding the filter outputs in phase opposition.

The implementation of the iterative DFT requires high-precision arithmetic because the poles are very close to the unit circle. Hence, it was necessary to do all calculations for this part of the system using double-precision arithmetic. Storage of the coefficients r, r^N, and W^{-k} in only 16-bit precision was permissible as long as care was taken to choose a value for r that gave an accurate representation of r in 16 bits. The value used for r was 0.998781, or 0.77730, octal.

IV. DERIVATION OF SPECTRAL ENVELOPE

A. CONVERSION TO POLAR COORDINATES

The next step, after obtaining the real and imaginary parts of 128 DFT coefficients spanning the region from 0 to π, is to convert to polar coordinates. For this purpose, an expanded version of an algorithm proposed by C. Rader (personal communication) was used. This method is particularly suitable to fixed-point single-precision arithmetic, and therefore will be described here.

The desire is to compute the magnitude r and the phase ϕ given the real and imaginary components x and y. The first step is to restrict the angle to $0° < \phi < 45°$, by redefining x and y such that $0 < y < x$, interchanging x and y if necessary, keeping a record of the original values for later adjustment of ϕ.

The method, as flowcharted in Fig. 5, involves iterative rotations of the vector $x + jy$ by an angle Θ which keeps decreasing in size as the vector converges on the real axis. Whenever the rotated vector is closer to the real axis than the original vector, the vector is replaced by the rotated vector and a record is kept of the rotation angle, as shown in the figure. If the new vector lies in the fourth quadrant $(y < 0)$, it is reflected about the x-axis to maintain a positive

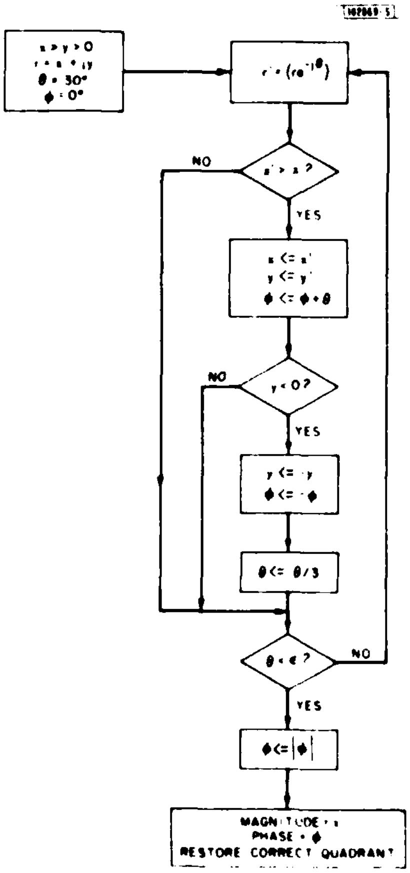

Fig. 5. Flowchart of algorithm used to derive magnitude and phase from real and imaginary components.

angle. Eventually, the y component becomes essentially zero, the x component is the same as the original r which was to be determined, and the accumulation of the rotations and inversions is the original phase angle ϕ. For each DFT coefficient, values for r (the magnitude spectral coefficient) and ϕ (the phase component) are computed in this manner.

B. SMOOTHING THE MAGNITUDE SPECTRUM

The magnitude spectrum can be viewed as the product of an excitation spectrum and a frequency-shaping spectrum. It is common in vocoder work, though not strictly correct, to incorporate the high-frequency falloff due to the fact that glottal pulses are not strictly impulses into the vocal-tract component, and thus to view the excitation spectrum as a flat sequence of equally spaced impulses at the harmonics of the fundamental frequency. We will take the same point of view toward separating the two components. Consequently, an appropriate method is to divide the magnitude spectrum by a smoothed version of itself, thus, in principle, leaving only the harmonic structure, while removing the formant structure.

The choice of an "optimal" smoothing filter is not immediately obvious, although some insight can be gained from previous work on related systems. Even the simple issue of whether or not the filter should be linear is not straightforward. For example, the cepstral method (homomorphic analysis) for extracting a spectral envelope has often been used in speech processing; this method is equivalent to smoothing linearly in the log magnitude domain.[12] Smoothing the magnitude-squared spectrum (power spectrum) is equivalent to low-time liftering of the autocorrelation function. Linear prediction[13] is also based on the autocorrelation function; its goal is to match exactly the first p autocorrelation coefficients, and force the remainder to fit a linear all-pole model.

Channel vocoders derive the spectral envelope from a set of bandpass filters whose bandwidths are considerably wider than those of the filters in the present system.[14] However, each appropriate wideband filter can be derived by adding a number of adjacent linearly weighted filter outputs, available as DFT coefficients. The set of narrowband-filter outputs can thus be reduced to a smaller set of wideband-filter outputs, spanning the frequency range. This smaller set is derived by smoothing in frequency the real and imaginary components of the original set of DFT coefficients. The magnitude output for each derived wider filter can then be computed from the corresponding x and y values.

However, the coherent combining of adjacent filters has effectively narrowed the speech time window. The magnitude, as a function of time, will tend to have peaks whenever the window is centered on a harmonic, and valleys when harmonics are near the edges of the window. Hence, it will be necessary to further lowpass-filter in time each magnitude function in order to remove this unwanted pitch information (pitch ripple). Such lowpass filtering is also necessary in channel vocoders.

A final method – one which has not been common in the vocoder world, but one which seems to represent a convergence point of all of the above methods – is to smooth in frequency the magnitude functions derived from the original set of filters. Smoothing the magnitudes rather than the real and imaginary components separately, as in a typical channel vocoder, achieves an incoherent rather than a coherent combination of the outputs of the various filters that are linearly weighted. In this sense, it resembles the cepstral and autocorrelation methods. By adding incoherently, the narrowing of the time window does not occur and, therefore, temporal pitch ripple is not a problem. Pitch ripple in the frequency domain can be removed if an adequate amount of frequency filtering is done.

Smoo'hing the magnitude can also be viewed as a compromise between the two extremes – smoothing the squared magnitude and smoothing the log magnitude. The former, corresponding to the autocorrelation method, expands the dynamic range prior to smoothing. The latter, the homomorphic method, reduces the dynamic range. Both these methods have accuracy problems in fixed-point arithmetic. They also require considerable computation for functions such as the log and square root.

Given all the above considerations, we felt that the best method to use in the context of the present system was to smooth in frequency the magnitude spectrum. If the cutoff time of the filter is below the first harmonic, the pitch ripple in frequency will be removed. The choice of cutoff time for the filter depends upon the fundamental frequency, and hence, ideally, should be adaptive. Too much smoothing results in a smeared spectral envelope; too little leaves undesirable pitch ripple in the spectrum.

Given that it has been decided to smooth the magnitude spectrum to derive the spectral envelope, the question of an appropriate smoothing filter remains unresolved. Here the familiar frequency-time trade-off problems are present in reverse. Whereas with filter design one is usually willing to allow poor temporal resolution in order to obtain a sharp cutoff frequency for a lowpass or bandpass filter, it is not clear that similar constraints are appropriate for the present situation.

The inverse transform of the magnitude spectrum is a function which resembles an autocorrelation function, as shown in Fig. 6(a-e). It might at first be expected that a filter with a sharp cutoff time just below the peak at the fundamental period (arrow in the figure) would be a good choice. However, such a choice would be equivalent to multiplying the waveform by a square window. The spectral resolution in the transform of a square-windowed segment is quite poor. The spectral coefficients, which are intended to describe a narrow frequency component of the transfer function of the system, are in fact including distal frequencies present in the tails of the sin nx/sin x filter. Hence, it may be preferable to taper the window in time in order to assure a more accurate frequency-domain characterization of the transfer function.

Given the above arguments, we decided to use a 17-point raised cosine as the lowpass filter for the spectrum, as shown in Fig. 7(a-b). This filter is quite narrow in frequency, i.e., has good frequency resolution. Its time-domain characteristics are given in Fig. 7(b). About 60 time samples are included in the central lobe, corresponding to a fundamental period of 3.6 ms. A fundamental period below this amount will result in the inclusion of some excitation information in the spectral envelope. Thus, this filter is adequate for male voices, but will leave some pitch ripple in female fundamental frequency ranges. We felt that additional smoothing would begin to remove significant portions of the spectral envelope information. This filter was found to be effective, judging by the quality of the output speech, even when the input was female speech, despite obvious pitch ripple still present in the spectrum.

In summary, the process for obtaining the spectral envelope by smoothing the magnitude spectrum is illustrated in Fig. 6(a-e). The example selected is a portion of the vowel in the word "old," spoken by a female. Figure 6(a) shows the 256-sample segment of the original waveform that was processed. The magnitude spectrum of the cosine-windowed waveform is shown in Fig. 6(b). This spectrum contains the harmonics of the fundamental frequency shaped with the spectral envelope. Figure 6(c) shows the inverse transform of the magnitude spectrum, as described previously, and Fig. 6(d) shows the portion of this signal that remains after processing the magnitude spectrum through the FIR filter. Although these two waveforms are never actually

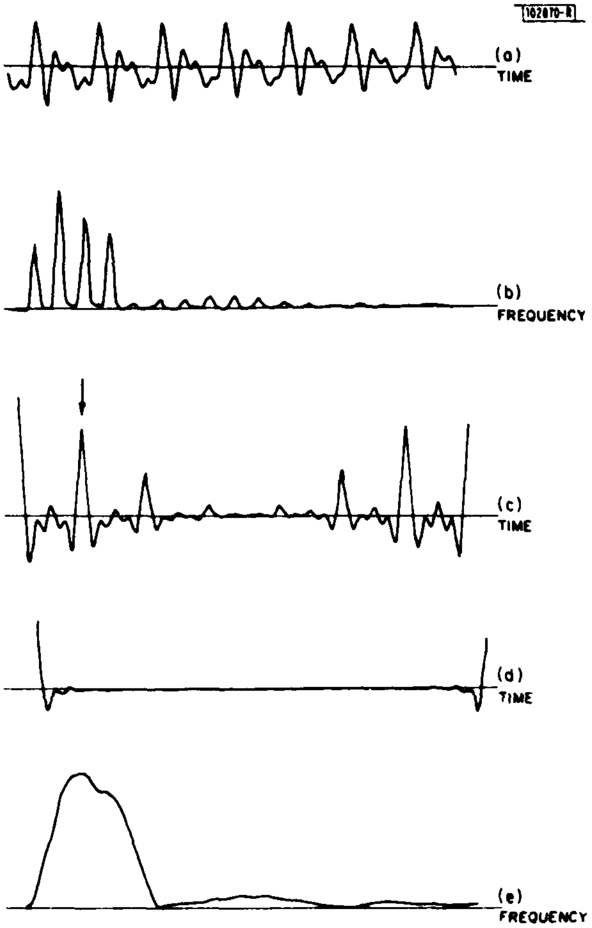

Fig. 6. Illustration of effect of filtering magnitude spectrum to derive spectral envelope. (a) Original waveform; (b) magnitude spectrum (including harmonic structure); (c) inverse transform of (b), arrow is at fundamental period; (d) result of multiplying (c) by time-domain characteristic of filter in Fig. 7; (e) spectral envelope derived by filtering using filter in Fig. 7.

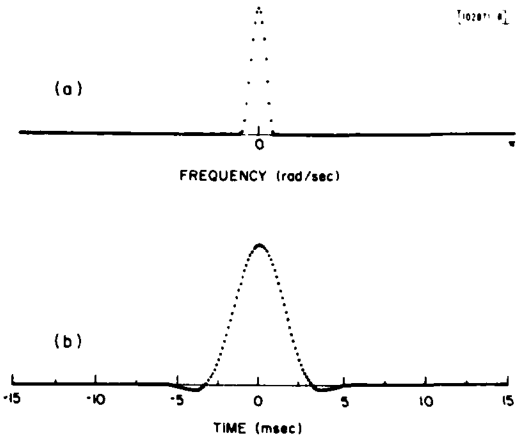

Fig. 7. Frequency-domain (a) and time-domain (b) characteristics of FIR filter used to smooth magnitude spectrum.

obtained by the system, they serve to illustrate, in the time domain, the effect of the smoothing filter. The waveform in Fig. 6(c) resembles, but has no simple relationship to, an autocorrelation function. The spectrum of this waveform is the square root of the spectrum of the autocorrelation of the original x(n) with itself. The portion of this waveform which describes the spectral envelope information is contained mainly before the peak identified by the arrow, located one fundamental period away from zero time.

Because the autocorrelation function is twice as long as the original waveform x(n), the power spectrum of x(n) corresponds to a time-aliased autocorrelation function. It is clear, however, that the energy in the waveform in Fig. 6(c) is already well attenuated by the time the sample at N/2 is reached. Therefore, the aliasing of data near zero time, due to the folding-in of data near N, will probably be negligible.

Figure 6(e) shows the spectrum that is obtained by smoothing the magnitude spectrum using the FIR filter shown in Fig. 7(a-b). This is also the DFT of the waveform shown in Fig. 6(d) where it is evident that the harmonic structure has been removed, but the spectral envelope is retained. This result is anticipated because the peak at the arrow in Fig. 6(c) is not present in Fig. 6(d).

Once the smoothed magnitude spectrum is obtained, a sample-by-sample divide achieves the deconvolution. Implicit in the process is the assumption that the vocal-tract resonance filter is a zero-phase filter. While this assumption is invalid, it has been frequently made for previous speech applications without noticeable degradation. Furthermore, it is not yet evident how to extract the proper vocal-tract phase information from the phase spectrum.

C. AN EXAMPLE: FREQUENCY COMPRESSION

Before proceeding to the more complicated issues of phase modification and high-frequency regeneration, it is worthwhile to describe in detail an application which does not require these additional steps.

A system of potential use for people suffering from high-frequency hearing loss is one which remaps the spectral envelope into the low-frequency region, while preserving the pitch. The reconstructed speech has the correct temporal characteristics and correct prosodics. The formant frequencies are shifted downward from their correct locations, and the bandwidths are

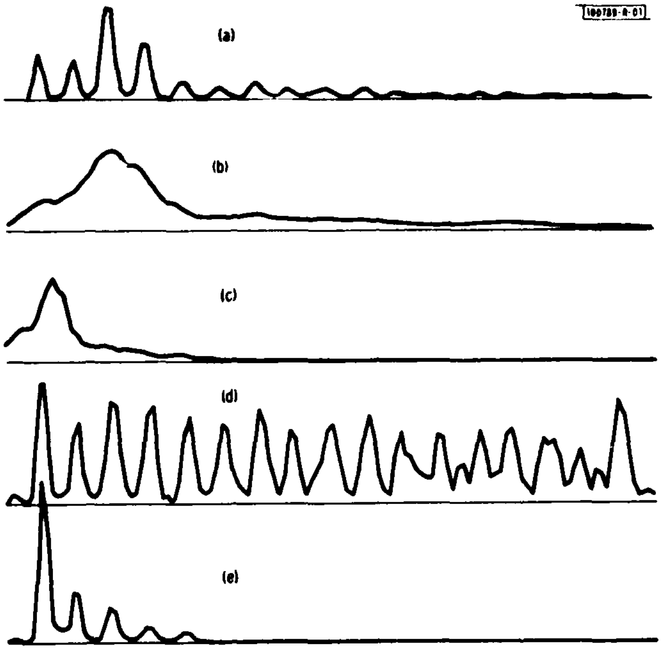

Fig. 8. Illustration of technique used to nonlinearly compress spectral envelope by overall factor of 3 while preserving fundamental. (a) Original magnitude spectrum, (b) spectral envelope, (c) remapped spectral envelope, (d) flattened spectrum, and (e) reshaped spectrum.

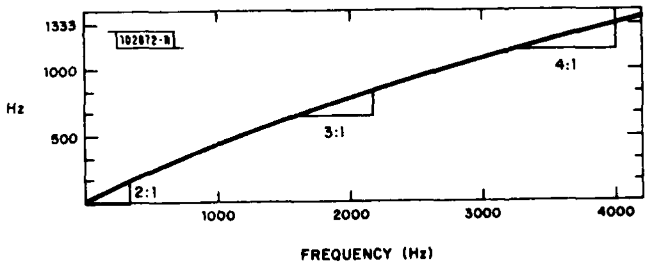

Fig. 9. Nonlinear compression scheme used for frequency-lowering experiment.

compressed accordingly. The advantage is in bringing the spectral energy of speech sounds such as /s/ and /t/ into the hearing range of the subject. While it has not yet been demonstrated that systems which attempt this sort of transformation are helpful, it is unclear that such systems have as yet been adequately designed or tested.[10]

The present system has tremendous inherent flexibility in the choice of a remapping scheme. If the fundamental frequency is to be preserved, all that must be done is to design a spectral remapping function appropriate for the person whose hearing is impaired. As an example, a scheme which compresses the spectrum overall by a factor of 3 has been implemented, as shown in Fig. 8(a-e). The selected compression scheme is nonlinear with only 2:1 compression at the low frequencies, and 4:1 compression at the highs, using the frequency-transformation function plotted in Fig. 9. Interpolated spectral samples are derived by linear interpolation between available samples. Since the spectrum is slowly varying in frequency, this choice is probably adequate. Ultimately, the 128 spectral samples are mapped into 43 samples, and multiplied by the first 43 excitation magnitude spectral samples to reshape. The upper two-thirds of the excitation spectrum is discarded, and the available 43 samples are added in phase opposition to obtain the output sample y(n).

Speech produced in this manner has well-preserved prosodic characteristics and is judged by informal listening tests to be of superior quality, in that there is very little noise superimposed on the signal and no perceived roughness or transient noises. However, because of the severely distorted spectrum, the speech is unintelligible to untrained ears. Extensive training tests will be necessary before the utility of the method can be properly evaluated.

V. HARMONIC REGENERATION FOR BASEBAND-EXCITED VOCODER

A. INTRODUCTION

Another potentially useful application for the system is as a baseband-excited vocoder. For this application, only a small fraction of the excitation spectrum is kept, from which the remainder of the excitation spectrum is reconstructed. The spectral envelope is downsampled, coded, and transmitted, along with the quantized baseband, and the regenerated excitation spectrum at the receiver is reshaped using the transmitted spectral coefficients.

For this application, an algorithm is needed for shifting the available harmonics by the correct amount such that the regenerated harmonics are at frequencies which are integer multiples of the fundamental. The method which has been devised to automatically make an intelligent choice as to the correct amount to shift is described below.

B. MECHANISM TO REDUCE FREQUENCY-OFFSET PROBLEMS

Suppose, for example, that the waveform to be processed is a sequence of pulses spaced by the fundamental period. This signal is lowpass-filtered at a cutoff frequency of $\pi/2$, 2:1-downsampled, and transmitted. The receiver must reconstruct an estimate of the upper band from the available low-frequency data.

At the receiver, the waveform is first 2:1-upsampled to restore the correct sampling rate. A DFT of the upsampled waveform will yield a spectrum which is correct only in the lower half of the range. From the lower-frequency spectral data, the upper half of the spectrum must be reconstructed.

Assume that the last harmonic available in the transmitted waveform is at ℓf_0, as shown in Fig. 10. Then the missing harmonics, which were present in the upper band of the original signal, are at $(\ell + 1) f_0$, $(\ell + 2) f_0$, etc. A frequency shift of the first half of the positive frequency spectrum by ℓf_0 would yield a magnitude spectrum identical to the discarded upper spectrum of the original signal. A reconstructed waveform which contains both the frequencies of the downsampled waveform and the frequencies of the waveform derived from the frequency-shifted spectrum would, except for possible phase shifts, be a replica of the original waveform.

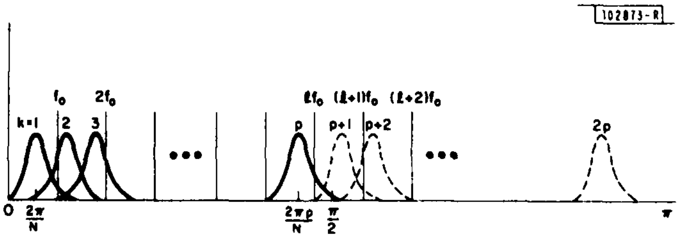

Fig. 10. Mechanism for generating upper half of excitation spectrum from available lower half. Input waveform is assumed to be a periodic pulse sequence. f_0, $2f_0$, etc. are frequencies of harmonics present. $2\pi k/N$ is center frequency of k^{th} filter. Entire available set of filters is frequency shifted by ℓf_0 to generate phantom set, indicated by dashed lines. Phantom harmonics are at correct frequencies, but phantom filter center frequencies are offset by an amount $\ell f_0 - 2\pi p/N$.

Thus, the synthesizer must shift the entire available positive-frequency spectrum by the frequency of the last available harmonic. Negative frequencies should be shifted by an equivalent negative harmonic frequency. One way to determine the center frequency of the last harmonic is to compute a DFT of the waveform and measure the location of the last peak. Parabolic interpolation between available DFT samples would probably be necessary to refine the location of the harmonic. A distinction would also have to be made between peaks which are true harmonics and spurious peaks which may be present in the spectrum. The method would thus be algorithmic, and the approach would resemble pitch detection.

Another approach, appealing because it is more automatic and insensitive to binary decision errors, is to use the phase information in the last correct DFT coefficient to determine the frequency of the last available harmonic. Each DFT coefficient, as a function of time, is the output of a narrow-bandpass frequency-sampling filter, as shown in Fig. 10. If the input spectrum is a set of harmonics spaced by a frequency width which is greater than the bandwidth of the filter, then, as a function of time, the output of the filter will be a sine wave at the frequency if_o of the harmonic passed by that filter:

$$y_k(n) = A_k(if_o)\ e^{jif_o n}$$

In particular, the last filter to pick up valid data is centered at a frequency slightly less than $\pi/2$. This filter's output, $y_p(n)$ in Fig. 10, is a sine wave at the frequency of the last available harmonic. Hence, as shown in the figure, this waveform has frequency ℓf_o and can be described by the equation

$$y_p(n) = A_p(\ell f_o)\ e^{j\ell f_o n}$$

where $A_p(\ell f_o)$ is the amplitude of the filter gain at the frequency ℓf_o of the ℓ^{th} harmonic, and n is a time index. Therefore, the phase of this output will be $(\ell f_o n)$ modulo 2π.

This phase is therefore the integral in time of the frequency of the ℓ^{th} harmonic. Ideally, if the filters are narrower than the harmonic spacing, each filter output will contain a phase component which is the integral of the frequency of the harmonic passed by that filter. A very simple mechanism for frequency-shifting the entire filter bank by the frequency of the last harmonic is to simply add the measured phase of the last filter's output to the measured phase of each filter output individually. The effect is to generate a set of phantom harmonics at frequencies $(\ell + 1)\, f_o$ through $(2\ell)\, f_o$ and a set of phantom filters passing these harmonics centered at frequencies $\ell f_o + 2\pi k/N$. The phantom harmonics will be at the correct locations, but the phantom filters in the upper-half band will be offset from a true set of 2p frequency-sampling filters (i.e., a DFT of the original waveform) by an amount equal to the offset of the ℓ^{th} harmonic from the center frequency of the p^{th} filter.

This technique, in theory at least, places the phantom harmonics at the correct frequencies. There is no need to even unwrap the measured phase to avoid discontinuities in 2π, since such discontinuities remain at 2π boundaries in the shifted harmonic frequencies. There are, however, two maladjustments of the phantom set relative to a true complete set of filters, i.e., DFT coefficients, characterizing the original impulse train.

If the phantom set of high-frequency filters is reshaped using transmitted spectral coefficients, these coefficients will be reintroduced at slightly erroneous frequencies. Since the spectral envelope changes slowly with frequency, this offset is probably imperceptible.

The other problem is that the first phantom filter is not spaced from the last true filter by the correct amount. If the last harmonic is below the center frequency of the last filter, then the first phantom filter will be too close to the last true filter. The result is that the energy in the reconstructed excitation is enhanced at the boundary. On the other hand, if the filter spacing is too great at the boundary, the excitation will have an apparent dip in energy in that frequency region. The effect of introducing a spectral peak or dip also should not be significant perceptually, since the excitation is rarely completely flat even in ideal conditions.

The logic behind the choice of this method for frequency shifting is that a major, perceptually significant, error (namely harmonic offset) is replaced by an equally severe mathematically, but presumably far less important perceptually, offset in the filter-bank frequencies. The method is dependent upon the assumption that filter bandwidths are narrow enough to pick up only a single harmonic most of the time, but not so narrow that no harmonics are present in the output. Experimentally, the problem of no harmonics seems to be more detrimental to quality than the problem of multiple harmonics in the present system. It is likely that wider filters or a more sophisticated algorithm, which may simply mean better precision in computation, may be beneficial in some cases for a high-pitched female-voice input.

This technique has been used for two applications of the system – the baseband-excited vocoder and the female-to-male voice-conversion algorithm. In both cases, a low-frequency band was copied up several times into the higher-frequency regions by adding a multiple of the phase measured from the last filter output. The baseband-excited vocoder is discussed in Sec. C below. However, the description of the voice modification will be deferred until Sec. VI, after time-scale modification has been addressed.

C. BASEBAND-EXCITED VOCODER

In the previous section, we described a mechanism for generating higher harmonics from lower ones, which will automatically duplicate them such that the upper harmonics will be multiples of the fundamental. Such a reconstruction has applications in the area of mid-range (10,000 to 20,000 bps) voice-encoding techniques, which will be addressed here.

The general class of baseband-excited vocoders, as shown in Fig. 11, models the spectral envelope by a set of parameters and transmits the baseband, usually as a lowpass but sometimes

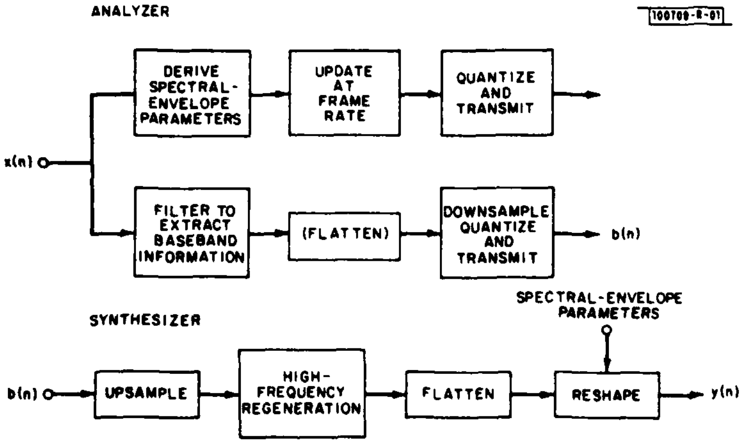
ANALYZER
100708-R-01
DERIVE SPECTRAL-ENVELOPE PARAMETERS
UPDATE AT FRAME RATE
QUANTIZE AND TRANSMIT
x(n)
FILTER TO EXTRACT BASEBAND INFORMATION
(FLATTEN)
DOWNSAMPLE QUANTIZE AND TRANSMIT
b(n)
SYNTHESIZER
SPECTRAL-ENVELOPE PARAMETERS
b(n)
UPSAMPLE
HIGH-FREQUENCY REGENERATION
FLATTEN
RESHAPE
y(n)

Fig. 11. General structure for baseband-excited vocoder.

as a bandpass signal containing only the information in the first formant region. This baseband is sometimes spectrally flattened at the analyzer prior to transmission, and at other times is transmitted intact and kept at the synthesizer to replace the portion of the spectrum that it contains. It ultimately must be modified in some way to reconstruct the missing harmonics, flattened, and then reshaped using the transmitted spectral-envelope parameters to recover the complete spectrum, as shown in Fig. 11.

The mechanism described in the previous section for generating the higher harmonics from the lower ones can be used in the context of essentially any baseband-excited vocoder to generate a complete excitation spectrum. The vocoder which was implemented, based on the overall structure of the present system, does not have a reasonable bit rate and is not separated into distinct analyzer and synthesizer components. However, it serves to illustrate an upper limit on the expected quality of a vocoder making use of this technique. Information content is reduced only in the sense that the upper three-quarters of the excitation spectrum, magnitude and phase, is discarded and replaced with an excitation derived from the lower one-quarter. The downsampling in frequency and time of the spectral-envelope spectrum, and quantizing of the baseband and spectral parameters, has not been attempted in the present context. However, potential systems that should yield a vocoder with an overall bit rate of around 10 kbps, and should be implementable in close to real time using presently available hardware, will be discussed.

A block diagram of the uncoded vocoder is given in Fig. 12. By lowpass-filtering in frequency the magnitude spectrum, as described in Sec. II, 128 spectral-envelope coefficients were obtained. The first 32 samples of the excitation magnitude spectrum were derived by dividing sample-by-sample in frequency the magnitude spectrum by the first 32 samples of the spectral envelope. The first 32 samples of the phase spectrum were copied up by adding multiples of the phase measured from the 32nd channel to all 32 to generate three copies spanning the region

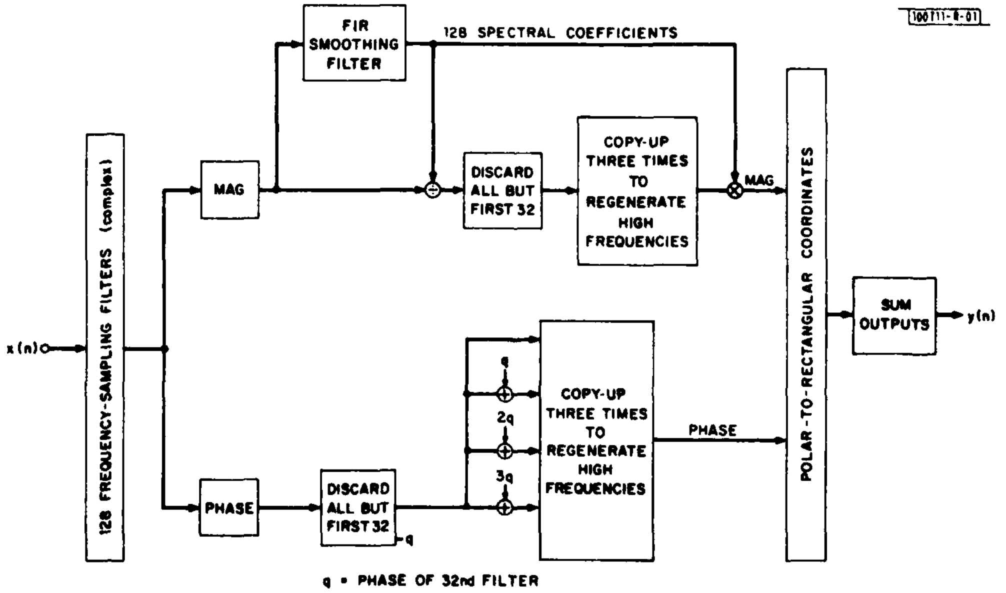
100711-R-01
FIR SMOOTHING FILTER
128 SPECTRAL COEFFICIENTS
128 FREQUENCY-SAMPLING FILTERS (complex)
MAG
DISCARD ALL BUT FIRST 32
COPY-UP THREE TIMES TO REGENERATE HIGH FREQUENCIES
MAG
POLAR-TO-RECTANGULAR COORDINATES
SUM OUTPUTS
y(n)
x(n)
q
2q
3q
COPY-UP THREE TIMES TO REGENERATE HIGH FREQUENCIES
PHASE
PHASE
DISCARD ALL BUT FIRST 32
q
q = PHASE OF 32nd FILTER

Fig. 12. System structure used for implemented uncoded baseband-excited vocoder. Analyzer and synthesizer are not separate components.

from 1000 to 4000 Hz. The magnitude spectrum was also duplicated three times, with no modifications. The resulting estimated excitation spectrum was then reshaped using the 128 spectral coefficients. Returning to rectangular coordinates and summing the outputs results in the output sample y(n). The process was repeated at the sampling rate.

The output was found to be of superior quality, nearly indistinguishable from the original speech in the case of male voices. Occasionally, for female voices, some distortion was audible, resembling the frequency-offset problems of vocoders which duplicate the spectrum by adding fixed-frequency offsets. It is likely that the problem is due to the fact that the 32nd filter is too narrow, such that the harmonic recovered is of very low amplitude. This, combined with the noise problems inherent in fixed-point arithmetic, generates a noisy phase function measured from the 32nd filter, thus introducing noise in the frequency offset.

It is likely that a set of wider filters would be advantageous for the case of female pitch ranges. When the phases of the i^{th} and the p^{th} filters are added, a correct result will be generated even if one of the two contains multiple harmonics, as long as the other one contains a single harmonic. Another possibility would be to examine the last several filters, and select the one with the largest amplitude to copy up. In this case, the phase would have to be unwrapped and a phase function for copy-up purposes generated by adding to the previous phase estimate the first difference of the phase measured out of the filter with the highest amplitude response. This mechanism is somewhat algorithmic, and therefore less attractive than the original method of simply adding the unwrapped phase.

In order to convert the above system into a functional vocoder, it must be split apart into an analyzer and a synthesizer, and downsampled estimates of the spectral envelope must be upsampled at the synthesizer and convolved with the copied-up and flattened baseband. A very simple analyzer structure is suggested in Fig. 13. An FFT is computed every 20 ms, and from the magnitude spectrum a spectral envelope is derived by smoothing. By choosing a $\pi/4$ cutoff time, the spectrum can be 4:1-downsampled in frequency without further loss of information or introduction of temporal aliasing. Only the upper three-quarters of this spectrum needs to be transmitted, and hence 24 spectral coefficients are sufficient. These can be quantized and transmitted once per frame. The baseband need not be flattened at the analyzer. Hence, a 1000-Hz cutoff lowpass filter of the original signal permits 4:1-downsampling to yield a baseband signal which can then be quantized by whatever method and transmitted.

At the receiver, as shown in Fig. 14, the baseband signal is first 4:1-upsampled with zero samples inserted between data samples. The upsampled signal is then processed through a 32-channel complex frequency-sampling filter bank spanning the first quarter of the spectrum. Dividing the derived magnitude spectrum by a smoothed version of itself completes the flattening process. The flattened magnitude spectrum and phase spectrum are duplicated three times, as shown in the figure, with additions of multiples of the phase measured from the 32nd filter, achieving the frequency shift as before. Next, the generated 128-point excitation spectrum is converted to rectangular coordinates. The real parts of the upper three-fourths of the excitation spectrum are linearly weighted to reduce the data to 24 coefficients which can then be reshaped with upsampled (in time) and lowpass-filtered spectral-envelope coefficients, as shown in Fig. 15. Finally, the real parts of the upper three-quarters of the derived spectrum are added to the real parts of the coefficients of the original 32 baseband frequency-sampling filter outputs, to regenerate the output sample y(n).

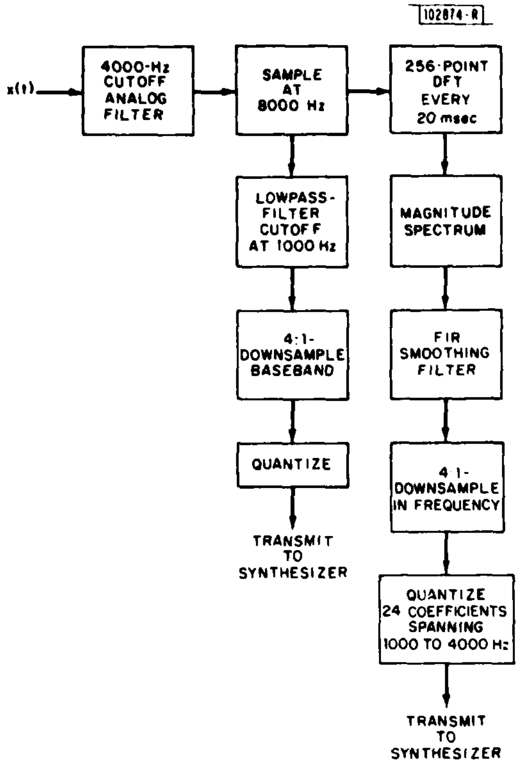
102874-R
x(t)
4000-Hz CUTOFF ANALOG FILTER
SAMPLE AT 8000 Hz
256-POINT DFT EVERY 20 msec
LOWPASS-FILTER CUTOFF AT 1000 Hz
MAGNITUDE SPECTRUM
4:1-DOWNSAMPLE BASEBAND
FIR SMOOTHING FILTER
QUANTIZE
4:1-DOWNSAMPLE IN FREQUENCY
TRANSMIT TO SYNTHESIZER
QUANTIZE 24 COEFFICIENTS SPANNING 1000 TO 4000 Hz
TRANSMIT TO SYNTHESIZER

Fig. 13. Proposed structure for analyzer of possible baseband-excited vocoder.

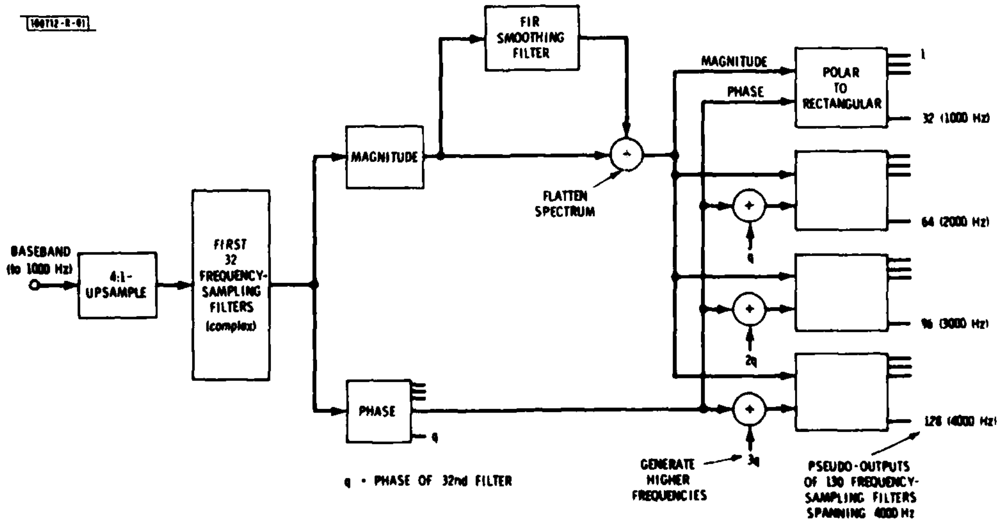

Fig. 14. Method for generating excitation spectrum from baseband.

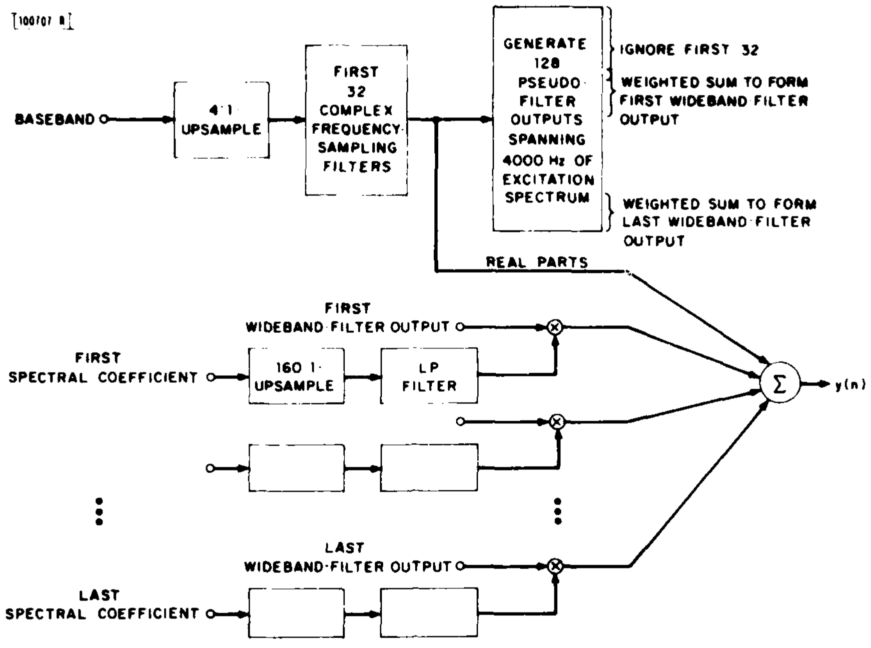

Fig. 15. Proposed structure for synthesizer of baseband-excited vocoder.

It is difficult to predict how much additional degradation will be incurred in the downsampling, upsampling, and quantizing processes. Of particular concern is whether the quantization of the baseband will result in the introduction of noise into the measurement of phase from the last filter, thus reducing the accuracy of the copy-up procedure.

VI. TIME-SCALE MODIFICATION

Since it is basically an extended phase vocoder, the system is capable of simply modifying temporal characteristics, using the same principles as a phase vocoder uses. The mechanism for modifying time can also be applied in an identical fashion for modifying the frequency scale. The D/A clock rate controls which dimension is altered. For some of the present applications, for example voice modification, both frequency and time are modified by the appropriate amounts.

In order to understand how time/frequency modification works, it is useful to consider an input which is a single sine wave. This sine wave is passed by the k^{th} filter, and conversion to magnitude and phase yields a constant magnitude and an unwrapped phase equal to the integral of the frequency:

$$\phi = 2\pi fn \quad .$$

If the unwrapped phase is multiplied by a factor c, the effective result is that the frequency of the sine wave is altered to cf. Depending upon the values of c and f, it may be necessary to increase the sampling rate in order to avoid frequency aliasing. For example, if f is near the half-sampling frequency and c is 2, the sampling rate must be doubled in order to accommodate a frequency of 2f. This upsampling can be accomplished by upsampling and lowpass-filtering the magnitude and unwrapped phase, as both of these functions would be slowly varying for a sufficiently narrow filter.

If the factor c is a fraction, the frequencies are all shifted down, so the resulting waveform will, in general, be oversampled. For most purposes such oversampling is of no consequence, but, if desired, the magnitude and phase functions could be upsampled, lowpass-filtered, and downsampled to the appropriate sampling frequency. If the factor c is 1/2, two-to-one downsampling in time of the output waveform with no filtering is permissible.

In general, a waveform is much more complicated than a single sine wave. Speech, at least in voiced regions, can be fairly well characterized as a sum of sine waves spaced by the fundamental frequency. If the filter bandwidths are sufficiently narrow that only one harmonic is picked up by each filter, then the above arguments apply.

There are several issues which become conceptually obscure as soon as the phase component of the DFT coefficients is altered, and these issues deserve some discussion at this point. When the phase is multiplied by a factor, not only the frequencies of the speech itself are being altered, but also the frequency and phase of the cosine window. For example, if c is 1/2, then the frequency of the cosine window is reduced by a factor of 2. Furthermore, the phase of the cosine window is also altered by a factor of 1/2, resulting in a conversion of the cosine into a sine, thus rotating the peak point of the window from the center point.

The main serious consequence of this window modification is that filters which formerly added in phase opposition may no longer do so. Consider an example of a single steady-state sine wave at the center frequency of the k^{th} filter, as the waveform to be processed. For this case, the cosine-windowed DFT coefficients are all zero except the values at $k - 1$, k, and $k + 1$, which will have coefficients $-1/4$, $+1/2$, and $-1/4$, respectively. The phase for all these three will be $2\pi kn/N$, equal to the integral of the frequency of the sine wave. However, in computing the magnitude and phase, the minus sign of the two side filters will be converted to a π offset in phase. When the phase is then multiplied by the factor 1/2, this π offset will be converted to $\pi/2$. When the resulting three filter outputs are added in phase opposition, the sum will be incorrect.

The problem is that, although the desire is only to change the frequencies of the generated sine waves, in the process their phase relationship relative to one another is also altered. The problem is only serious because a given sine wave is picked up by multiple filters. In order to obtain the correct energy of the component, it is necessary to add these outputs in the proper phase relationship.

A solution is to remove the π phase opposition relationship prior to phase multiplication. This step can be accomplished by adding π to the phase of alternate filters, effectively rotating the time window by N/2 samples. Hence, the desired sample of the inverse DFT is the first one rather than the middle one. Now, considering the case of a single sine wave, the coefficients for the three adjacent filters will be +1/4, +1/2, and +1/4, and so the phase measured by all three will be identical. Thus, by simply adding the outputs after phase modification, the correct amplitude of the altered sine wave will be obtained.

The problem of phase unwrapping must also be resolved, and again the π discontinuities enter in. In the present system, phase was unwrapped in time only. Since phase is available at the sampling rate, the procedure is quite straightforward. 2π discontinuities are easy to recognize and remove. A good algorithm is to project the previous estimate of the phase by an increment $2\pi k/N$, where k is the center frequency of the filter, and to add multiples of 2π to the current measured value of the phase until it is within π of the estimate.

Whenever the frequency of the harmonic passed by the filter crosses the boundary from the central lobe to one of the negative side lobes, π discontinuities occur. It is not correct to remove the π discontinuity, as this would result in guaranteed improper phase relationships among adjacent filters.

The desire is to measure the frequency of the output waveform picked up by the filter and to alter that frequency by a factor c without affecting the phase relationship between the output of this filter and that of adjacent filters. A method of achieving this goal is to force the unwrapped phase to be within π of the projected phase estimate, but to keep a record of whether an even or an odd number of 2π increments was necessary. The measured frequency f ($\nabla\phi$ after unwrapping) can be multiplied by the factor c, and then an estimate of the modified phase can be obtained by adding cf to the former estimate if an even number of increments was necessary, or adding cf + π if an odd number was necessary.

By thus separating out the π discontinuities from the frequency estimate, it becomes more feasible to upsample and lowpass-filter the frequency estimate in time in order to generate extra samples in the case when a spreading of the frequency axis is desired. Otherwise, the π discontinuities would be smoothed out and the result would be incorrect.

Since only a single sample of the inverse DFT is desired, it is permissible to downsample in frequency the outputs of the modified filter bank prior to adding them, by discarding the information in every other filter. To see that this is so, consider the matrix in Fig. 16(a-b) of the coefficients for each cosine-window filter output. By adding the outputs of adjacent filters, a coefficient of one is retrieved for each of the members of the original filter set, prior to the introduction of the hanning window. If alternate filters are discarded, then each filter of the original set will have associated with it a coefficient of 1/2, alternately contributed by a single filter or by two adjacent filters in the downsampled set.

Another way to view the downsample is to consider it as temporal aliasing of the cosine window. The central part of the window is folded into the tails, and the zero sample of the inverse

(a) ADD FULL SET

[102875-5]

	x_1	x_2	x_3						x_{n-1}	x_n
$\hat{x}_n$	1/4								1/4	1/2
$\hat{x}_1$	1/2	1/4								1/4
$\hat{x}_2$	1/4	1/2	1/4							
$\hat{x}_3$		1/4	1/2	1/4						
$\hat{x}_4$			1/4	1/2	1/4					
$\hat{x}_5$				1/4	1/2	1/4				
$\hat{x}_6$					1/4	1/2	1/4			
$\hat{x}_7$						1/4	1/2	1/4		
$\sum_i \hat{x}_i$	1	1	1	1	1	1				

(b) ADD DOWNSAMPLED SET

$\hat{x}_n$	1/4								1/4	1/2
$\hat{x}_2$	1/4	1/2	1/4							
$\hat{x}_4$			1/4	1/2	1/4					
$\hat{x}_6$					1/4	1/2	1/4			
$\sum_{i\ \text{even}} \hat{x}_i$	1/2	1/2	1/2	1/2	1/2	1/2				

Fig. 16(a-b). Matrix of hanning-window spectral coefficients illustrating why 2:1-downsampling of windowed spectrum is permissible. X(k) are samples of square-windowed spectrum. $\hat{X}(k)$ are samples of hanning-windowed spectrum.

DFT of the downsampled set is the sum of the contribution at the tip of the tail (i.e., 0) and the contribution at the center of the window. Hence, the aliased zero sample is the correct output.

The downsample is an advantageous step because it considerably reduces the amount of frequency overlap between adjacent filters. Hence, even if the phase relationships among adjacent filters are wrong, the error in the computed combined magnitudes will be greatly reduced. It also, of course, reduces the amount of computation necessary.

To summarize, in order to change temporal characteristics the appropriate procedure is as follows. First, only 64 of the DFT coefficients are kept, once the raised cosine window has been introduced. These are a downsampled set of the positive frequency data. The information in these 64 coefficients is converted to polar coordinates, and the phase in each is unwrapped in time only. The first difference of the unwrapped phase is multiplied by a factor c, which is less than one if the desire is to speed up the speech. If c is greater than one, then upsampling and lowpass-filtering in time of the magnitude and frequency (first difference of the unwrapped phase) of each DFT coefficient is necessary at this point.

Finally, the modified phase is recovered from the frequency estimates, the polar coordinates are converted back to rectangular coordinates, and the 64 real parts are summed to form the output sample y(n). If the speech is to be slowed down, interpolated spectra are also converted back to rectangular coordinates, and the outputs summed to form interpolated samples of y(n). The whole procedure is repeated at the sampling rate.

VII. VOICE MODIFICATION

The present system has the capability of performing voice modification by simultaneously modifying the excitation and the spectral envelope. Such transformations have potential applications in the areas of speaker normalization for computer speech recognition, improvements in quality of vocoded speech, studies of male-female voice-quality differences, and psychological experiments concerning sex-role models.

One approach to the problem of speaker normalization for speech recognition is to convert all voices into a canonic voice prior to the main processing needed to identify the spoken utterance. The system would tune to the new speaker by computing an average spectral envelope and average fundamental frequency from a few selected sentences. From these average values would be derived a (probably nonlinear) spectral remapping function, which could include a spectral tilt correction component as well, and a factor by which to adjust the fundamental frequency. Each utterance spoken by the speaker would then be transformed by a normalization process derived from the averaged data from the test set.

It is generally the case that parametric vocoders such as LPC and channel vocoders produce quality which is speaker-dependent. The type of speaker-normalization approach discussed above could potentially be applied to reduce the sensitivity of vocoder quality to speaker differences. As a particular example, it is generally recognized that parametric vocoders generate inferior-quality synthetic speech when the input is female speech. It may be possible to improve the quality of vocoded female speech by first processing the waveform through a voice-transformation system. The most important modification would probably be to transform the fundamental frequency into a male range. The vocoder parameters obtained from the transformed speech should characterize the spectrum more accurately, because the spectrum is more finely sampled by the harmonics. In addition, the pitch extractor should perform better in the male-range fundamental region. The correct fundamental could be restored in the synthesizer by simply multiplying the extracted fundamental frequency by the correct inverse transformation factor.

Another potentially useful application for voice-conversion schemes is to aid in a study of male/female voice-quality differences. After the obvious spectral- and fundamental-frequency differences have been removed, remaining differences such as breathiness become perceptually evident. Further changes, such as a time-varying spectral-remapping function determined by the spoken speech sound, or addition of noise to simulate breathiness, might further reduce subtle differences. Eventually, perhaps, a male voice could be converted to a voice which perceptually was identical to a female voice. In the process of tuning the system, the perceptually significant differences in the two voice types would be identified.

A fourth application is in the area of sex-role-model studies. If the transformation process is successful enough that the listener can be fooled, then interesting studies could be carried out examining the different reactions to the same passage spoken by a female as contrasted with a male. Intonation, speaking style, and spoken content would all be identical; only the sex of the speaker would differ. Thus, the reaction to the sex of the speaker can be isolated from the reactions to other subtle differences in manner of speaking.

In order to convert a male voice into a female-like voice, it is necessary to change the fundamental frequency and the formant frequencies by different amounts. Experimentally, it has been shown that female formants on the average are about 20-percent higher than male formants.[15]

Thus, a simple 20-percent expansion of the spectral envelope is a good first approximation to the correction for vocal-tract length. Male and female mean fundamental frequencies, in general, differ by substantially more than 20 percent. Therefore, the change in fundamental frequency must be done independently. A change in fundamental frequency by a factor of 2 is usually enough to bring a male fundamental into the female range.

The spectral envelope incorporates the glottal pulse frequency shaping into the transfer function. Hence, the high-frequency falloff due to the glottis will be altered by the same mechanism which alters the formant frequencies. The fundamental frequency will be modified by remapping the frequency scale in the excitation spectrum. To properly simulate sex conversion, data should be collected on glottal falloff for females and for males, and spectral tilt factors introduced to account for the measured differences.

Since the female vocal tract is not simply a linearly scaled version of the male vocal tract (the pharynx is shorter by a greater amount than is the mouth),[16] it may be anticipated that a simple linear-scale correction for the spectral envelope is not adequate. However, the relationship between vocal-tract length and formant frequencies is not simple, and, in fact, there is evidence that women articulate differently from men in order to compensate for the differing length ratios of the mouth and laryngeal cavities. Goldstein[9] recently proposed that linearly scaled female formant frequencies differ from male frequencies only in that the vowel space is expanded outward, i.e., females enunciate more clearly. In the context of the present system, a linear scale might therefore produce a male who seems to overarticulate or a female who seems to speak too casually.

For present purposes, a simple 20-percent expansion (male to female) or compression (female to male) of the spectrum was done along with an approximately 2:1 adjustment of fundamental frequency. The purpose is simply to demonstrate the quality of the reconstructed speech when significant alterations of both spectrum and fundamental frequency are attempted. The quality of the modified voice and the intelligibility of the reconstructed speech are good measures of the success of the system in the deconvolution, reconvolution process.

To convert a male voice into a female-like voice, the phase spectrum was multiplied by a factor > 1 to effectively raise the fundamental frequency. The spectral envelope was interpolated to obtain a new set of coefficients such that the first 4800 Hz of the expanded excitation spectrum could be reshaped with a 20-percent expanded spectral envelope. The resulting magnitude and phase components contain information up to 4800 Hz, and thus at the original 8-kHz sampling rate would be undersampled. A solution would be to upsample and lowpass-filter the magnitude and phase components prior to the return to real and imaginary parts. A simpler solution is to stay with the 8-kHz sampling rate but discard samples above 4000 Hz in the reshaped excitation spectrum.

The mechanism used for the conversion scheme is outlined in Fig. 17. The fundamental frequency is modified by a factor of 1.8, by multiplying every sample of the unwrapped phase spectrum by this amount. Next, the spectral envelope is linearly interpolated to map 128 samples into 86 samples (128 * 1.2/1.8). This step achieves both the 20-percent spectral expansion and the adjustment to allow sample-by-sample multiplication by the modified excitation spectrum. Finally, the excitation spectrum is reshaped, and only the first 72 samples, extending up to 4000 Hz, are kept for the final sum to restore y(n).

It may be recalled in the discussion of time-scale modification that a partial amelioration of the phase-error problem was realized by downsampling the filter set prior to the summation to

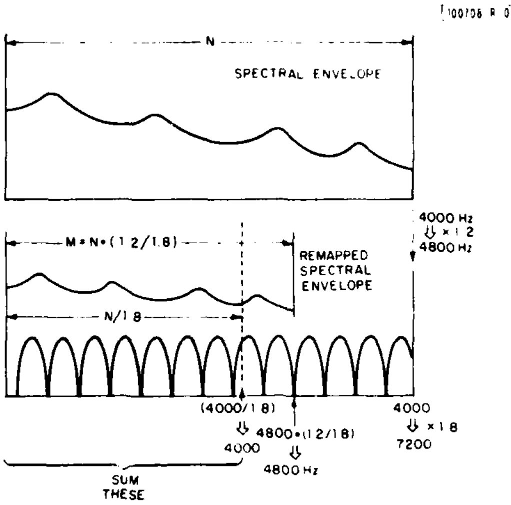

Fig. 17. Schematic illustration of mechanism used to alter fundamental frequency by factor of 1.8 and frequency-expand spectral envelope by 20 percent, to convert a male voice to a female-like voice. Phase spectrum is first multiplied by 1.8 to remap frequency scale. Spectral envelope is resampled from *N* to *M* samples prior to reshaping first 4800 Hz of modified excitation spectrum. Only first 4000 Hz of reshaped spectrum are kept, to avoid frequency aliasing.

form y(n). The downsampled set has the advantage that frequency overlap is considerably reduced and, therefore, incoherent recombination is not so severe a problem. However, when the spectral envelope is restored in altered form, such downsampling can no longer be done with impunity, as we can no longer guarantee no temporal aliasing. Particularly in the case of converting a male voice to a female-like voice when more spectral information is crammed into less frequency space, it is possible that an additional 2:1-downsampling, reducing the spectrum to less than 42 samples, may be denying an adequate description of the spectral envelope.

For the male-to-female conversion process, a 17-point raised cosine was used as the FIR filter to smooth the magnitude spectrum to obtain the spectral envelope. The first zero point of the inverse transform of this filter is at approximately $\pi/4$, and therefore little aliasing will occur with a 4:1-downsample of the spectral envelope. Keeping only 42 samples is only a 3:1-downsample, and hence should be permissible to avoid significant temporal aliasing of the spectral envelope.

For conversion of a female voice to a male-like voice, an analogous inverse process ensues, as shown in Fig. 18. In this case, extra harmonics must be generated from the available ones. A complete duplication of the excitation spectrum generates more than a sufficient number of harmonics. However, since the harmonic structure of the upper half of a female spectrum tends to be somewhat irregular, we decided that a better method for generating a more regular excitation spectrum would be to copy up the first half of the excitation spectrum four times. The copy-up procedure is as described in Sec. VI, with multiples of the phase measured in the filter at center frequency $\pi/4$ being added to the phase of the entire set for each duplication.

The unwrapped phase of the resulting duplicated spectrum is multiplied by the factor 0.6 in order to compress the harmonic spacing by this amount, as shown in Fig. 18. Then, the spectral

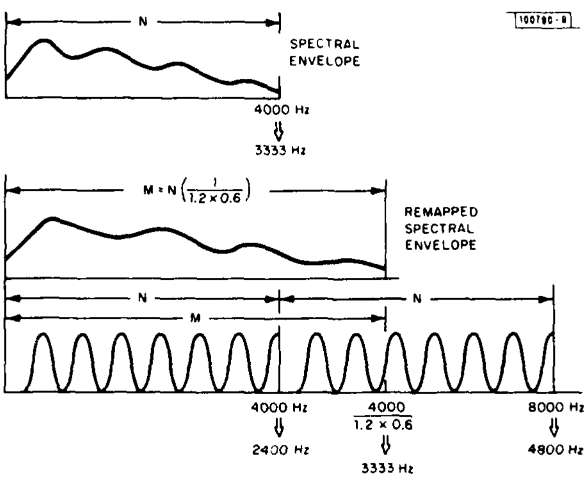
N
SPECTRAL ENVELOPE
4000 Hz
⇓
3333 Hz
$M = N\left(\frac{1}{1.2 \times 0.6}\right)$
REMAPPED SPECTRAL ENVELOPE
N
N
M
4000 Hz
⇓
2400 Hz
$\frac{4000}{1.2 \times 0.6}$
⇓
3333 Hz
8000 Hz
⇓
4800 Hz

Fig. 18. Schematic illustration of mechanism used to alter fundamental frequency by factor of 0.6 and compress spectral envelope by 20 percent, to convert a female voice to a male-like voice. Excitation spectrum is first duplicated to generate additional harmonics. Then, phase spectrum is multiplied by 0.6 to redefine frequency axis. Spectral envelope is resampled from N to M samples prior to reshaping first 3333 Hz of modified excitation spectrum. Remainder of excitation spectrum is discarded.

envelope is remapped from 128 samples to 178 [128/(1.2 * 0.6)] samples. Samples between the available 128 samples of the spectral envelope are generated by linear interpolation between adjacent available samples. Since the spectral envelope is oversampled by nearly a factor of 4, we felt that linear interpolation was adequate to recover intermediate coefficient estimates. These 178 samples are then post-multiplied by the corresponding first 178 samples of the duplicated excitation magnitude spectrum. These samples span a frequency scale of only 3333 Hz in the modified domain. Spectral information beyond this frequency is unavailable due to the original 8-kHz sampling rate. If it is desired to include frequencies up to 4 kHz in the output speech, it is necessary to record the input at a 20-percent increased sampling rate.

VIII. RELATIONSHIP TO FFT PHASE VOCODER

A. INTRODUCTION

The system described here requires a somewhat exorbitant amount of computer time. Even using a high-speed microprocessor, the processing is about 30 times real time. Thus, with presently available digital hardware, a real-time version is not feasible. Therefore, it is clearly desirable to consider other methods of implementing the system that might be computationally more efficient, but qualitatively equivalent.

The most promising alternate method is to implement the phase vocoder using FFT techniques for both the analysis and synthesis. The current system obtains a DFT at the sampling rate, but since each DFT coefficient is the output of a narrow-bandpass filter, this sample-by-sample update is clearly a grossly oversampled system.

Implementation of the analyzer using FFT techniques is very straightforward. However, not until very recently was an FFT method for the synthesizer also worked out. Portnoff's doctoral thesis describes such a method.[4] The key mathematical steps involve showing that upsampling and smoothing the spectral coefficients prior to computing the inverse DFT (IDFT) (a single sample only) is mathematically equivalent to upsampling and interpolating samples, properly chosen, of the waveforms derived from the inverse DFTs, computed at the downsampled rate.

B. REVIEW OF FFT PHASE VOCODER

Imagine that a DFT of a set of N windowed samples of the input is available every R samples. These DFT coefficients can be viewed as a function of time as the R:1-downsampled outputs of a bank of narrow-bandpass filters. If each DFT filter output is multiplied by exp $[-j2\pi Rik/N]$, its output will be frequency-shifted down to the origin, and therefore will be a lowpass signal. It is then theoretically possible to upsample and lowpass-filter the downsampled-filter outputs to recover intermediate samples of the lowpass waveforms. If these intermediate samples are then frequency-shifted back up to the appropriate center frequency, a set of DFT coefficients can be recovered, at the sampling rate, from the undersampled set. As in the present system, the output samples of y(n) can be recovered by adding the outputs of the filters, now available at the sampling rate, in phase opposition. Mathematically:

$$y(m) = \frac{1}{N} \sum_{k=0}^{N-1} e^{j\pi k} e^{j2\pi mk/N} \sum_{i=-\infty}^{\infty} h(m - Ri)\, \tilde{X}_k(Ri)$$

where h(m) is an FIR lowpass filter, $\tilde{X}_k(Ri)$ are the DFT coefficients after the frequency shift to zero frequency, the term $e^{j2\pi mk/N}$ is the modulation back up to the original center frequency, and the $e^{j\pi k}$ is the phase opposition term to recover the middle sample of the window.

The two summation indices can be interchanged, after which it becomes apparent that the inner sum is in the form of a DFT:

$$y(m) = \sum_{i=-\infty}^{\infty} h(m - Ri) \frac{1}{N} \sum_{k=0}^{N-1} e^{j\pi k} e^{j2\pi mk/N} \tilde{X}_k(Ri) \quad .$$

Hence, if we let

$$\tilde{x}_{Ri}(m) = \frac{1}{N} \sum_{k=0}^{N-1} e^{j\pi k} e^{j2\pi mk/N} \tilde{X}_k(Ri)$$

then,

$$y(m) = \sum_{i=-\infty}^{\infty} h(m - Ri)\, \tilde{x}_{Ri}(m) \quad .$$

The time index m can be expressed as Ri + ℓ and, hence, $\tilde{x}_{Ri}(m)$ is the ℓ^{th}, or $[(m)]_R^{th}$ sample of the inverse DFT, where $[(m)]_R$ is m modulo R.

In the Portnoff implementation, all the IDFT buffers necessary for the lowpass-filter convolution with h(m) were kept and each y(m) was determined after all the necessary data had been accumulated. Hence, the system required a large amount of buffering which would have prohibited implementation on a small high-speed microprocessor.

However, both Holtzman[17] and Crochiere[18] have shown recently that the y(m) can be computed as partial sums updated when each new set of N samples of the inverse transform is computed. This mechanism greatly reduces the storage requirements and makes feasible implementation on a small computer.

The frequency shift down to zero frequency of individual filters effectively removes the linear phase component in the iterative DFT, thus resulting in a very slow phase change as a function of both frequency and time. If the DFT is updated sufficiently often, phase can be unwrapped relative to 2π discontinuities, although π discontinuities are more intractable than in the iterative system.

When the factor c is introduced, the frequency axis is effectively remapped, as before. However, the sampling rate is now controlled by the spacing between samples of the IDFT of the modified spectrum. Therefore, the sampling rate has also been altered by the factor c.

For example, if c is 0.5, then a 0- to 4000-Hz spectrum is remapped into 0 to 2000 Hz. An N-point inverse DFT of the remapped spectrum will generate N samples of an output waveform at an implicit 2:1-downsampled rate relative to the input-signal sampling rate.

Hence, in units of time, the window is now twice as long as the original time window. The overlap between frames is thus twice as great as in the original x(n) and, therefore, one advances by only R/2 samples (cR) in the output waveform y(n) for each advance of R samples in the original waveform. When the final output is played out at the original clock rate, frequencies are shifted back up to the original values and the speech is time-compressed by a factor of 2.

When the phase vocoder was implemented using the iterative DFT procedure, the spectrum was downsampled in frequency by a factor of 2 prior to the summation to form y(m). The downsampling reduced the amount of overlap between adjacent filters and therefore helped to ameliorate the consequences of phase errors. In the FFT version, if the same raised cosine window is used, it should also be possible to downsample the spectrum, since again only the middle sample of the IDFT is needed. The fact that the filtering is done after rather than before the computation of the IDFT is of no consequence, since the system is linear. The effect of the downsample, in the temporal domain, is to produce periodic copies of the windowed $\hat{x}_{Ri}(m)$

twice as often, which overlap one another by 50 percent. These aliased outputs can be multiplied by the corresponding lowpass-filter coefficients just as before, to yield the final outputs.

Mathematically:

$$y(m) = \frac{1}{N} \sum_{k=0}^{(N/2)-1} \exp\left[\frac{j2\pi mk}{N/2}\right] \sum_{i=-\infty}^{\infty} h(m - Ri)\, \tilde{X}_{2k}(Ri)$$

$$= \sum_{i=-\infty}^{\infty} h(m - Ri)\, \frac{1}{N} \sum_{k=0}^{(N/2)-1} \exp\left[\frac{j2\pi mk}{N/2}\right] \tilde{X}_{2k}(Ri)$$

$$= \sum_{i=-\infty}^{\infty} h(m - Ri)\, \hat{x}_{Ri}(m)$$

where

$$\hat{x}_{Ri}(m) = \frac{1}{N} \sum_{k=0}^{(N/2)-1} \exp\left[\frac{j2\pi mk}{N/2}\right] \tilde{X}_{2k}(Ri) \quad .$$

C. FFT PHASE VOCODER IN CONTEXT

In the previous section, we discussed how the FFT phase vocoder can be related to the filter-bank phase vocoder as implemented in the present system. Here, the issue of whether the modifications of excitation and spectrum can be performed in the context of the FFT phase vocoder is discussed. The major concern is whether the modifications will introduce wider filter frequency characteristics such that the intermediate values of the modified spectrum cannot be derived from the available ones.

Consider first the simple case of frequency compression with pitch preserved. In this case, the phase information is left unmodified and, therefore, unwrapping is not necessary. For each sample at which a spectrum is available, the spectral-envelope estimate can be obtained as before, by smoothing the magnitude spectrum. The sample-by-sample divide to deconvolve, and the sample-by-sample multiplication of the first third of the excitation spectrum by a downsampled-envelope estimate to reshape, yield the frequency-compressed spectrum consisting of only N/3 samples. This spectrum can then be zero-padded to N/2, the nearest power of 2, and then an inverse DFT will yield a 2:1-downsampled waveform $\hat{x}_{Ri}(m)$. The set of $\hat{x}_{Ri}(m)$ can then be convolved with the 2:1-downsampled lowpass filter h(m). Half as many output samples as input samples will be generated, and these can be played out at half the sampling rate, thus preserving time.

The modifications to the magnitude spectrum can be interpreted as a filtering of the waveform with a time-varying filter. The variations in the filter are controlled by the speed of movement of the articulators. In the context that a low-bit-rate vocoder succeeds in downsampling in time the spectral information, it can be assumed that such downsampling will be permissible here. A vocoder with a 20-msec update rate tends to suffer from blurring of the crispness of stops and other sounds where rapid changes are taking place. An FFT phase vocoder generally has to update much more frequently, on the order of every 5 msec, and therefore would probably be sampling the spectral-envelope information sufficiently often to obtain a good characterization of the changes as a function of time.

The next application to be considered is the conversion of a male voice into a female-like voice. In this case, phase must be unwrapped. Several of the upper harmonics are discarded, and the spectral envelope is reintroduced in the lower portion of the excitation spectrum. Again, the spectrum can be zero-padded to the nearest power of 2 – in the case of our example, padded back to N samples. The factor c by which the phase was multiplied controls the highest frequency present (center frequency of the last DFT coefficient in the zero-padded spectrum), which, in turn, controls the sampling rate of the $\hat{x}_{Ri}(m)$ computed from the inverse DFT.

The system can be viewed as a time-expansion phase-vocoder system, on which has been superimposed a time-varying filter corresponding to the spectral-envelope reshaping. The time-expanded output is played back at c times the original clock rate in order to restore time, and expand frequencies. Again, there are no conceptual difficulties in the implementation.

Systems which need to copy up the available harmonics in order to generate additional high-frequency harmonics are those which are most uncertain as to the quality of the synthesized speech generated using the FFT phase-vocoder version. These include the baseband-excited vocoder and the female-to-male-like voice transformation. Adding the phase measured out of the last available DFT coefficient to the phases measured from all the others can be done, as before, but it will generate outputs which are less constrained as to frequency content. The effective bandwidth of each pseudofilter in the higher band will be equal to the sum of the bandwidths of the two filters that were combined to generate it. Hence, the spectrum must be updated twice as often in order to be able to recover the intermediate samples.

Consider a simple case of copying up the excitation once and introducing the spectral envelope spread over the double spectral range. If the phase information is then multiplied by the factor 0.5, the result is a lowering of the fundamental frequency by a factor of 2 while preserving the spectrum. For this example, there are 2N DFT coefficients spanning only the original 4000-Hz band. The inverse DFT yields a waveform $\hat{x}_{Ri}(m)$ spanning a 2N sample-window duration, at the original sampling rate. In other words, the hanning window is now twice as long in absolute time, but the sampling rate is identical to that of the input waveform x(n). The set of $\hat{x}_{Ri}(m)$ can be filtered using the same lowpass-filter coefficients as were used for the case of no time change, to generate an output waveform with both time and spectral envelope preserved, but fundamental frequency lowered by a factor of 2.

For the more complicated case in which the spectrum is also altered by 20 percent, the only difference is that the spectral-envelope information no longer fills out the entire 2N-point spectrum. Zero-padding to 2N will recover a system much like the one above, and the procedure from that point is identical.

Thus, there are no conceptual difficulties in realizing any of the above transformations using an FFT phase vocoder. The issues which need to be investigated concern mainly how often the DFT must be computed to get superior quality in terms of the spectral-envelope modeling, and in terms of the phase unwrapping, and whether π discontinuities in phase will introduce adverse effects. It remains to be demonstrated, by actually implementing an FFT version, whether the theoretical treatment has overlooked some inherent flaw.

IX. RESULTS

Six versions of the system were implemented on the Lincoln Digital Speech Processor. These versions differ from one another only in the way in which the excitation spectrum and spectral envelope are recombined to form the output waveform y(n). These six versions perform the following six transformations:

(a) 2:1 time expansion of the waveform; spectrum and fundamental unaltered.

(b) 2:1 time compression of the waveform; spectrum and fundamental unaltered.

(c) Nonlinear 3:1 compression of the spectral envelope; fundamental-frequency and temporal characteristics preserved.

(d) Male → Female-like: Temporal characteristics preserved; fundamental frequency changed by a factor of 1.8; spectrum expanded by 20 percent.

(e) Female → Male-like: Temporal characteristics preserved; fundamental frequency changed by a factor of 0.6; spectrum compressed by 20 percent.

(f) Uncoded vocoder: Temporal, spectral, and fundamental-frequency characteristics unaltered; first 1000 Hz of excitation used to regenerate upper 3000 Hz.

Twelve 2-sec sentences, six male and six female, were processed through each of the systems, although only half the utterances were processed through systems (d) and (e) (voice modification). The processed signals were selectively subjected to spectrographic and spectral analyses, and the quality of the output speech was judged by informal listening tests; no intelligibility tests were performed on the data.

Figure 19 (a-f) shows a sequence of waveforms and spectra for a 32-msec portion of the vowel in the word "great" spoken by a male. In the figure, the waveform (a) and spectrum (b) of the original utterance are compared with the waveforms and spectra of the outputs of the male-to-female conversion system and the 3:1 frequency-compression system.

The waveform of the female-like utterance is shown in Fig. 19(c), and its spectrum is in Fig. 19(d). In comparing these with the original, it is evident that excitation pulses are now

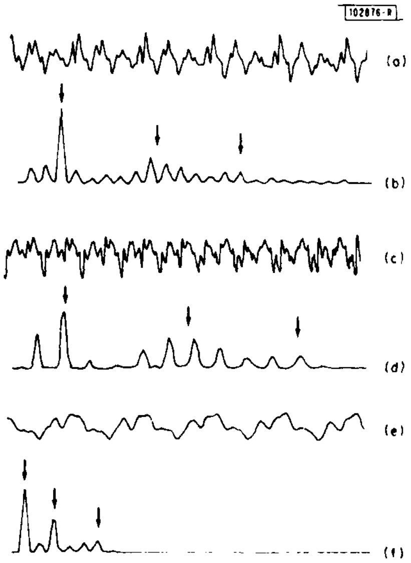

Fig. 19. Waveforms and spectra for original speech and transformed speech using two transformation systems. (a) 31-msec window of waveform in middle of diphthong /e/ in word "great" spoken by a male; (b) spectrum of waveform in (a); (c) waveform derived by conversion of (a) to female-like speech; (d) spectrum of waveform in (c); (e) waveform derived by 3:1 frequency-compressing spectral envelope of (a); and (f) spectrum corresponding to waveform in (e).

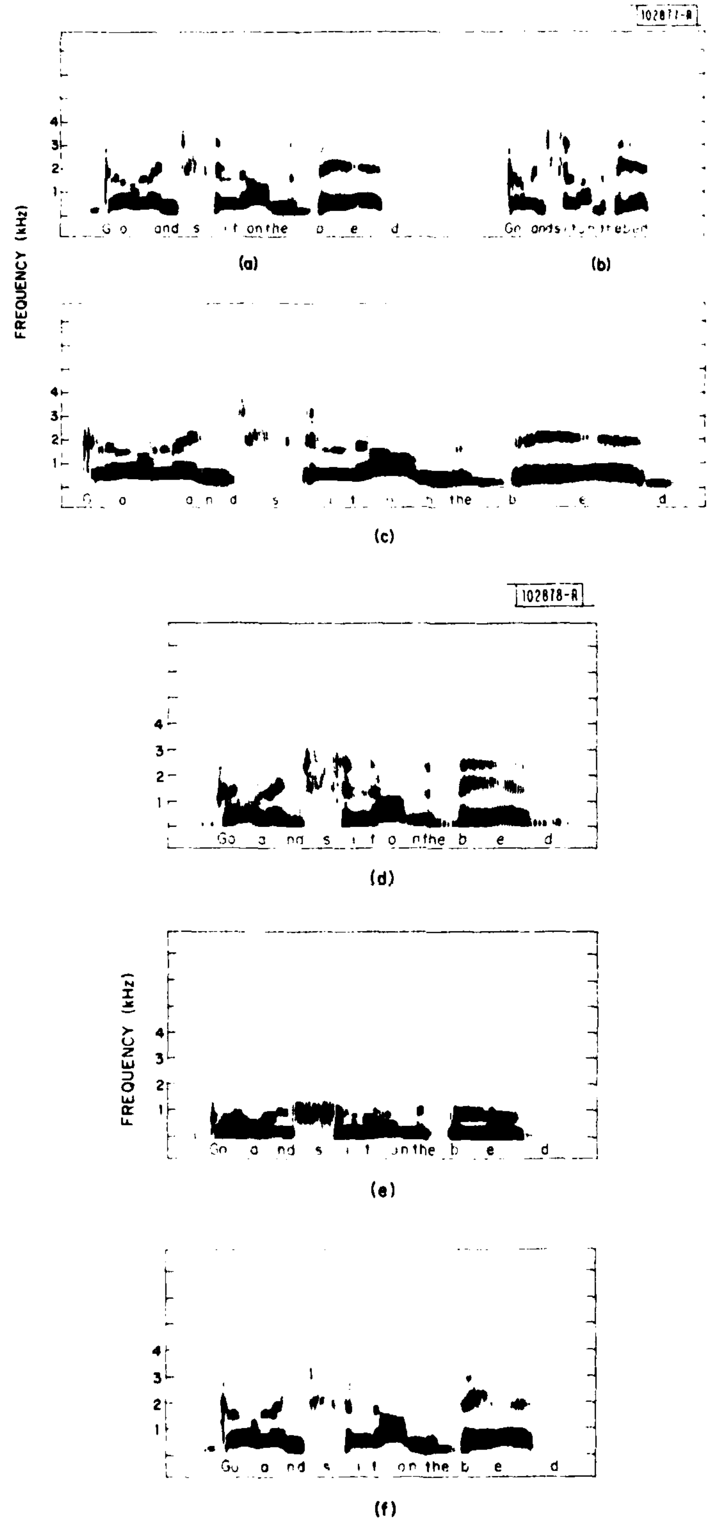

Fig. 20. Spectrograms of original sentence and outputs of five transformation systems. (a) Original sentence, "Go and sit on the bed.", spoken by a female; (b) 2:1 time compressed; (c) 2:1 time expanded; (d) converted to male-like voice; (e) 3:1 frequency compressed; and (f) 1000-Hz baseband-excited vocoder.

occurring more frequently in time, and harmonics are correspondingly more widely spaced in frequency. While the spectral envelope still exhibits three resonances corresponding to the first three formants (shown by the arrows), these have been displaced upward in frequency by about 20 percent.

Figure 19(e) shows the waveform of the 3:1 frequency-compressed output. This waveform clearly contains only low-frequency information, but the spacing of the excitation pulses is the same as in the original utterance. The spectrum of this waveform, shown in Fig. 19(f), has essentially no energy above $\pi/3$ (1333 Hz). However, the three formant resonances are all present below this frequency.

A set of spectrograms was obtained for the output waveforms produced by five of the six systems when the input was the utterance "Go and sit on the bed," spoken by a female. A companion set was also obtained for the utterance "The goose laid an odd egg," spoken by a male. The set of spectrograms for the first utterance is shown in Figs. 20(a) through (f). The preservation of detail in frequency is quite exact for the 2:1 time expansion [Fig. 20(c)], although time compression [Fig. 20(b)] seems to have distorted spectral detail in some portions of the utterance – for example, the first formant at the end of the utterance. The pitch striations (vertical lines in the spectrum) for both these systems appear to be regular and relatively noise-free.

For the conversion to the male-like voice [Fig. 20(d)], the formants are lowered by 20 percent. The pitch striations are clearly spaced by a greater amount than in the original utterance, reflecting the lowering of the fundamental frequency. The formant bandwidths are perhaps somewhat irregular compared with a typical male utterance (notice the second formant in the word "bed"); such irregularity is probably a consequence of the inadequate sampling of the spectrum by the harmonics in the original female utterance.

The resolution of the formant structure in the 3:1 frequency-compressed utterance [Fig. 20(e)] is not adequate using the filters of a voiceprint machine. Hence, formants (particularly F1) appear to be constant over long intervals. However, it is clear that little information exists above about 1300 Hz, and that the formants and frication energy above 1300 Hz have been compressed rather than discarded.

In comparing Fig. 20(f) with Fig. 20(a), it is clear that the vocoder has restored the upper formants at essentially the correct frequencies, although formant amplitudes are sometimes somewhat attenuated relative to the original. The pitch striations, however, are quite accurate and smooth.

Judging by informal listening tests, the quality of the synthetic speech produced by all these transformations was quite good, although the frequency-compressed utterance is unintelligible to untrained ears. It is likely that an improvement in the intelligibility of the frequency-compressed utterances would be achieved if the frequencies below 200 Hz were attenuated. The high-energy content of the compressed first formant may be masking important information in the compressed second- and third-formant regions.[19]

The female-to-male-like conversion system produced an utterance which was highly intelligible and reasonably noise-free, but did not sound quite like a male. It is likely that overarticulation and breathiness are causing a retention of a female-like quality. It is also likely that male and female intonation patterns are different, a problem that would be difficult to adjust for, given the present system.

One can conjecture that possible losses in intelligibility and quality of the synthetic speech may be due to two broad categories of effects: computational degradation and inadequate modeling of the transformation process. For all the systems, quantization problems due to fixed-point arithmetic may be contributing to some low-level background noise. Systems which must scale the spectrum by modifying the phase component are susceptible to unwrapping errors in the phase plus the possibility of multiple harmonics appearing in a single-filter output. In the latter case, a simple doubling of the phase component will generate extraneous frequency components at incorrect frequencies. The same problems will also affect systems in which the higher harmonics are to be generated from the low-frequency band, such as the female-to-male-like conversion scheme and the vocoder.

The category of inadequate modeling includes assumptions that are made about how to characterize the excitation and the transfer function, and assumptions that are made concerning the transformation processes. The transfer function is assumed to have zero phase, an assumption that is known to be wrong. Furthermore, it was necessary to remove the high time components of the transfer function in order to remove the fundamental. Thus, the derived smooth spectral envelope may have formant bandwidths that are too broad compared with the true transfer function of the vocal tract. This smoothing could also tend to reduce the crispness of sounds such as stop bursts. Temporal smoothing may also reduce stop crispness.

All the systems implemented were simple examples of what can be done using the basic building blocks of the system. For example, the time-expansion system expands all segments of time by the same amount. If the goal is to emulate a person speaking slowly, then it is likely that the time expansion should be done nonlinearly, such that most of the expansion is taken up in steady-state vowel and consonant portions, whereas consonant releases should have about the same time course as in the original speech. Although the system can handle time-varying time change, an algorithm would be necessary to decide when to expand by what amount. Similar arguments apply for time compression.

Further complexities that should be introduced in some of the other systems have already been discussed. More careful modeling of the voice transformations is needed, such as further separation of the transfer function into glottal effects and vocal-tract effects, and independent adjustments of each of these components. Psychoacoustical issues should be addressed more carefully in the design of the frequency-compression system.

X. CONCLUSIONS

A speech analysis-synthesis system has been developed which is capable of linear scaling of the fundamental frequency and nonlinear or linear scaling of the spectral envelope so as to realize a large number of transformations on the speech waveform. The desired remapping of the spectrum can be specified explicitly, and depends upon the particular application. The system is impervious to voicing errors or errors in the measurement of the fundamental frequency, because the excitation function is never explicitly modeled. Furthermore, because the excitation function is obtained by deconvolving with the transfer-function filter, the speech is more natural sounding than speech generated using a periodic pulse train or random noise source as excitation.

The system is also capable of time-scale modification in conjunction with the transformations of the spectrum or fundamental frequency. This feature evolves because a phase vocoder is embedded in the system structure.

A version of the system operates as a baseband-excited vocoder. Higher harmonics are generated from the low-frequency band by frequency-shifting the available spectrum by an amount equal to integer multiples of the frequency of the highest harmonic present in the baseband. Again, explicit extraction of the value of this frequency is never done, so that transients which might arise due to decision errors do not occur.

Several aspects of this system can be pursued in greater depth. Since the system requires a considerable amount of processing time, a configuration using an FFT-type phase-vocoder structure would be essential before the system is practical. Using presently available hardware, the FFT version should be implementable in real time.

The baseband-excited vocoder which was implemented was only a demonstration system. Data reduction occurred only in that the excitation spectrum was reconstructed from the 0- to 1000-Hz frequency band. For conversion of that system into a practical baseband-excited vocoder, the system would have to be split apart into a separate analyzer and synthesizer, and the spectral envelope and baseband would have to be downsampled and encoded.

The frequency-compression scheme was also only a demonstration system. A more useful version of that system should run in real time and allow independent user adjustment of the amount to compress the spectrum and the amount to lower the fundamental frequency. Furthermore, the compression scheme for the spectrum was not carefully chosen based on psychoacoustical data. It may result that a simple linear-compression scheme yields more intelligible speech than the nonlinear scheme that was implemented. Furthermore, it may be necessary to preemphasize the compressed speech to reduce the masking effect of the first formant on the higher formant information.[19] Licklider and Miller[20] report only a 10-percent reduction in articulation score for speech which is highpass-filtered to remove all frequencies below 1000 Hz. Therefore, attenuation of those low frequencies should not affect overall intelligibility very much. Finally, a formal test involving extensive training of both unimpaired and hearing-impaired listeners on the frequency-transformed speech is necessary before an evaluation of the effectiveness of the transformation process in improving intelligibility is possible.

A version of the voice-transformation scheme may be useful at the front end of a speech-recognition system. In that case, additional processing would be needed to measure the mean fundamental frequency and mean formant frequencies of the voice. The measured values could then be used to determine the amount by which to adjust the excitation and spectrum to convert the voice into a canonic speaker for whom the system has stored templates.

The voice-transformation system could also be used to quantify the perceptually significant differences between male and female voices. For such a study, it may be necessary to use a time-varying frequency-remapping function which is adapted to the spoken speech sound.[16] Furthermore, a simple linear scaling of the excitation may not be adequate. For example, female voices often sound breathy compared with male voices. Such breathiness is attributed to a partially open glottis, and can be modeled by a glottal spectrum with greater attenuation in the high frequencies and a turbulence noise source superimposed on the periodic source.[21]

In summary, a speech analysis-synthesis system has been developed and implemented which exhibits considerable promise as a potential tool in several speech-related experiments. The system as implemented is not very practical because of its heavy computational requirements. However, the system should be able to be reconfigured at great savings in computation with little degradation in quality, using an FFT algorithm. Specific transformations have applications in the fields of speech bit-rate reduction, aids to the handicapped, psychological and psychoacoustical experiments, and speech-recognition systems.

REFERENCES

1. L. R. Rabiner and R. W. Schafer, Digital Processing of Speech Signals (Prentice-Hall, Englewood Cliffs, New Jersey, 1978).

2. L. D. Braida, N. I. Durlach, R. P. Lippmann, B. L. Hicks, W. M. Rabinowitz, and C. M. Reed, "Matching Speech to Residual Auditory Function - A Review of Past Research," ASHA Monograph (1978).

3. V. W. Zue, "Translation of Divers' Speech Using Digital Frequency Warping," Quarterly Progress Report 101, Research Laboratory of Electronics, M.I.T. (15 April 1971), pp. 175-182.

4. M. R. Portnoff, "Implementation of the Digital Phase Vocoder Using the Fast Fourier Transform," IEEE Trans. Acoust., Speech, and Signal Processing ASSP-24, 243-248 (1976).

5. J. L. Flanagan and R. M. Golden, "Phase Vocoder," Bell Syst. Tech. J. 45, 1493-1509 (1966).

6. B. Gold and J. Tierney, "Digitized Voice-Excited Vocoder for Telephone-Quality Inputs, Using Bandpass Sampling of the Baseband Signal," J. Acoust. Soc. Am. 37, 753-754 (1965), DDC AD-619923.

7. J. Makhoul and M. Berouti, "High Frequency Regeneration in Speech Coding Systems," Intl. Conf. on Acoustics, Speech, and Signal Processing, Washington, D.C., 2-4 April 1979, pp. 428-431.

8. C. J. Weinstein, "A Linear Prediction Vocoder with Voice Excitation," Proc. EASCON '75, Washington, D.C., 29 September - 1 October 1975, pp. 30A-30G.

9. U. Goldstein, "An Articulatory Model of the Vocal Tracts of Growing Children," PhD. Thesis, Department of Electrical Engineering, M.I.T. (June 1980).

10. P. E. Blankenship and V. J. Sferrino, "Succinct LDSP User's Guide," private communication (10 June 1977).

11. A. V. Oppenheim and R. W. Schafer, Digital Signal Processing (Prentice-Hall, Englewood Cliffs, New Jersey, 1975), pp. 160-163.

12. A. V. Oppenheim, "A Speech Analysis-Synthesis System Based on Homomorphic Filtering," J. Acoust. Soc. Am. 45, 458-465 (1969), DDC AD-689574.

13. B. S. Atal and S. L. Hanauer, "Speech Analysis and Synthesis by Linear Prediction of the Speech Wave," J. Acoust. Soc. Am. 50, 637-655 (1971).

14. B. Gold and C. M. Rader, "The Channel Vocoder," IEEE Trans. Audio Electroacoust. AU-15, 148 (1967), DDC AD-679147.

15. G. E. Peterson and H. L. Barney, "Control Methods Used in a Study of the Vowels," J. Acoust. Soc. Am. 24, 175-184 (1952).

16. G. Fant, "A Note on Vocal Tract Size Factors and Non-Uniform F-Pattern Scalings," Quarterly Progress and Status Report, Speech Transmission Laboratory (15 January 1967), pp. 22-30.

17. S. Holtzman, "A Statistical Measure for Feature-dependent Speed Transformations of Speech," ScM. Thesis, Department of Electrical Engineering, M.I.T. (June 1980).

18. R. E. Crochiere, "A Weighted Overlap-Add Method of Short-Time Fourier Analysis/Synthesis," IEEE Trans. Acoust., Speech, and Signal Processing ASSP-28, 99-101 (1980).

19. R. L. Wegel and C. E. Lane, "The Auditory Masking of One Pure Tone by Another and Its Probable Relation to the Dynamics of the Inner Ear," Phys. Rev. 23, 266-285 (1924).

20. J. C. R. Licklider and G. A. Miller, "The Perception of Speech," in Handbook of Experimental Psychology, S. S. Stevens, Ed. (Wiley, New York, 1951), p. 1052.

21. K. N. Stevens, "Physics of Laryngeal Behavior and Larynx Modes," Phonetica 34, 264-279 (1977).

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (*When Data Entered*)

19 REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER 18 ESD-TR-80-90	2. GOVT ACCESSION NO. AD-A097091	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (*and Subtitle*) 6 Speech Transformation System (Spectrum and/or Excitation) Without Pitch Extraction 9		5. TYPE OF REPORT & PERIOD COVERED Technical Report
		6. PERFORMING ORG. REPORT NUMBER Technical Report 541
7. AUTHOR(s) 10 Stephanie Seneff 12 50		8. CONTRACT OR GRANT NUMBER(s) 15 F19628-80-C-0002 ARPA Order-3673
9. PERFORMING ORGANIZATION NAME AND ADDRESS Lincoln Laboratory, M.I.T. P.O. Box 73 Lexington, MA 02173 14 TR-541		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS ARPA Order 3673 Program Element Nos. 61101E and 62708E Project Nos. 0D10 and 0T10
11. CONTROLLING OFFICE NAME AND ADDRESS Defense Advanced Research Projects Agency 1400 Wilson Boulevard Arlington, VA 22209 11		12. REPORT DATE 31 July 1980
		13. NUMBER OF PAGES 56
14. MONITORING AGENCY NAME & ADDRESS (*if different from Controlling Office*) Electronic Systems Division Hanscom AFB Bedford, MA 01731		15. SECURITY CLASS. (*of this report*) Unclassified
		15a. DECLASSIFICATION DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (*of this Report*)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (*of the abstract entered in Block 20, if different from Report*)

18. SUPPLEMENTARY NOTES

None

19. KEY WORDS (*Continue on reverse side if necessary and identify by block number*)

speech transformation	speech analysis-synthesis
baseband-excited vocoder	spectral envelope

20. ABSTRACT (*Continue on reverse side if necessary and identify by block number*)

A new speech analysis-synthesis system has been developed which is capable of independent manipulation of the fundamental frequency and spectral envelope of a speech waveform. The system deconvolves the original speech with the spectral-envelope estimate to obtain a model for the excitation. Hence, explicit pitch extraction is not required. As a consequence, the transformed speech is more natural sounding than would be the case if the excitation were modeled as a sequence of pulses. The system has applications in the areas of voice modification, baseband-excited vocoders, time-scale modification, and frequency compression as an aid to the partially deaf.